NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

ANALYTICALLY QUANTIFYING GAINS IN THE TEST AND EVALUATION PROCESS THROUGH CAPABILITIES-BASED ANALYSIS

by

Eric J. Lednicky

September 2011

Thesis Advisor:	Rachel T. Silvestrini
Second Reader:	Harrison C. Schramm

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE				*Form Approved OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** September 2011	**3. REPORT TYPE AND DATES COVERED** Master's Thesis
4. TITLE AND SUBTITLE Analytically Quantifying Gains in the Test and Evaluation Process through Capabilities-Based Analysis		**5. FUNDING NUMBERS**
6. AUTHOR(S) Eric J. Lednicky		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000		**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) Department of the Navy		**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol Number: N/A

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited	**12b. DISTRIBUTION CODE**

13. ABSTRACT (maximum 200 words)

Military operating environments are increasingly diverse and technically challenging. Fielding relevant weapons systems to meet the demands of this environment is increasingly difficult, prompting policy shifts that mandate a focus on systems capable of combating a wide threat range. The Capabilities-Based Test and Evaluation (CBT&E) construct is the Department of the Navy's effort to concentrate on integrated system design with the objective of satisfying a particular operational response (capability) under a robust range of operating conditions. One aspect of CBT&E is the increased employment of advanced mathematical and statistical techniques in the Test and Evaluation (T&E) process. This study illustrates advantages of incorporating these invaluable techniques, like Design of Experiments (DOE) and Modeling and Simulation (M&S), within the T&E process. We also suggest a general methodology for approaching test plan design, presented via a notional scenario in which a complex system must defend a forward outpost. We found through statistical analysis that the application of DOE concepts to the System Under Test (SUT) throughout three primary phases of T&E quantifiably improved the accomplishment of the selected Measure of Effectiveness (MOE).

14. SUBJECT TERMS Test and Evaluation, Capabilities-Based Test and Evaluation (CBT&E), Design of Experiments, Modeling and Simulation, Mission-Based Test and Evaluation, Operational Test & Evaluation (OT&E), Developmental Test and Evaluation (DT&E), Integrated Test and Evaluation (IT&E), SASIO, Defense Acquisition, Factorial Design, Optimal Design			**15. NUMBER OF PAGES** 107
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

ANALYTICALLY QUANTIFYING GAINS IN THE TEST AND EVALUATION PROCESS THROUGH CAPABILITIES-BASED ANALYSIS

Eric J. Lednicky
Commander, United States Navy
B.S., United States Naval Academy, 1994

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2011

Author: Eric J. Lednicky

Approved by: Rachel T. Silvestrini
Thesis Advisor

Harrison C. Schramm
Second Reader

Robert F. Dell
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Military operating environments are increasingly diverse and technically challenging. Fielding relevant weapons systems to meet the demands of this environment is increasingly difficult, prompting policy shifts that mandate a focus on systems capable of combating a wide threat range. The Capabilities-Based Test and Evaluation (CBT&E) construct is the Department of the Navy's effort to concentrate on integrated system design with the objective of satisfying a particular operational response (capability) under a robust range of operating conditions. One aspect of CBT&E is the increased employment of advanced mathematical and statistical techniques in the Test and Evaluation (T&E) process. This study illustrates advantages of incorporating these invaluable techniques, like Design of Experiments (DOE) and Modeling and Simulation (M&S), within the T&E process. We also suggest a general methodology for approaching test plan design, presented via a notional scenario in which a complex system must defend a forward outpost. We found through statistical analysis that the application of DOE concepts to the System Under Test (SUT) throughout three primary phases of T&E quantifiably improved the accomplishment of the selected Measure of Effectiveness (MOE).

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ANOVA	Analysis of Variance
AOI	Area of Interest
AT&L	Acquisition, Test, and Logistics
ATEC	Army Test and Evaluation Command
BAMS	Broad Area Maritime Surveillance
C3I	Command, Control and Communications
CBA	Capability Based Assessment
CBT&E	Capabilities-Based Test and Evaluation
CBTEWG	Capabilities Based Test and Evaluation Working Group
CNO	Chief of Naval Operations
COI	Critical Operational Issues
COMTEVOPFOR	Commander, Operational Test and Evaluation Force
COTS	Commercial Off-The-Shelf
DoD	Department of Defense
DOE	Design of Experiments
DON	Department of the Navy
DOT&E	Director, Operational Test & Evaluation Command
DSB	Defense Science Board
DT	Developmental Test
DT&E	Developmental Test & Evaluation
FOB	Forward Operating Base
FoS	Family of Systems
GLM	Generalized Linear Model
HMMWV	High Mobility Multipurpose Wheeled Vehicle
IED	Improvised Explosive Device
IOT&E	Initial Operational Test and Evaluation
ISR	Intelligence, Surveillance and Reconnaissance
IT	Integrated Test
IWC	Integrated Warfighting Capabilities
KPH	Kilometers per Hour
KM	Kilometers
LVC	Live, Virtual, and Constructive

M&S	Modeling and Simulation
MDAP	Major Defense Acquisition Program
MOA	Memorandum of Agreement
MOE	Measure of Effectiveness
MOP	Measure of Performance
MRAP	Mine Resistant Ambush Protected
NAE	Naval Aviation Enterprise
NAVAIR	Naval Air Systems Command
NAVSEA	Naval Sea Systems Command
NCIP	Naval Aviation Enterprise Capabilities Based Assessment Integrated Process
NPS	Naval Postgraduate School
OE	Operational Effectiveness
OFAT	One Factor at a Time
OPNAVINST	Operational Navy Instruction
OS	Operational Suitability
OSD	Office of the Secretary of Defense
OT	Operational Test
OT&E	Operational Test and Evaluation
OTA	Operational Test Agency
QRF	Quick Reaction Force
SASIO	Situational Awareness for Surveillance and Interdiction Operations
SBT&E	Specifications-Based Test & Evaluation
SME	Subject Matter Expert
SoS	System of Systems
SUT	System Under Test
T&E	Test and Evaluation
TEMP	Test and Evaluation Master Plan
UAS	Unmanned Aerial System
UAV	Unmanned Aerial Vehicle
UCLASS	Unmanned Carrier Launched Airborne Surveillance and Strike system
USAF	United States Air Force
USMC	United States Marine Corps
USN	United States Navy
USSOCOM	United States Special Operations Command
VPN	Virtual Private Network

EXECUTIVE SUMMARY

"I ... believe we have a belabored operational test and evaluation regime that from time to time, more often tends not to be able to deliver the integrated and the interoperable systems that we're going to need."

– Chief of Naval Operations, 19 Aug 2011, Association for Unmanned Vehicle Systems International conference (Weisgerber, 2011)

Increasingly complex military operating environments have strained the ability of current acquisition processes to field weapon systems that keep pace with technological advances. Traditional Test and Evaluation (T&E) methods narrowly focus on system design to satisfy a particular requirement or performance property, especially in the early developmental phases of design. We refer to this type of testing as Specifications-Based Test and Evaluation (SBT&E). This limits the ability of modern, complex systems to satisfy the capability requirements of the 21st century battlespace, primarily because the emergence of asymmetric threats has driven the Services towards a greater level of interoperability (Defense Science Board Task Force, 2008). The use of SBT&E under these conditions has contributed to the failure of many new systems during the operational test phase.

Service leadership has recognized and is increasingly concerned by the costly trend of increasing failures. To improve the T&E process, the Department of the Navy is shifting from SBT&E to a Capabilities-Based Test and Evaluation (CBT&E) process. CBT&E integrates the tactical employment of the prospective system into the design at the very earliest stages; this approach of 'beginning with the end in mind' has strong potential to lower both acquisition costs and time-to-deploy, resulting in more capability sooner to the field. Furthermore, CBT&E encompasses a broad focus on system design in order to satisfy a particular operational effect spanning the breadth of all phases of T&E. This ensures that the acquisition process delivers operationally effective systems relevant to a wide range of threats and passing the operational test phase.

The CBT&E process emphasizes the design of families of systems, which integrate individual capabilities to obtain a more capable “meta-system” greater than the sum of the individual parts, to meet military operational commitments (Popper, 2004). Additionally, CBT&E incorporates advanced scientific and statistical methods, such as Design of Experiments (DOE) and Modeling and Simulation (M&S) techniques, throughout the design process. The intelligent application of DOE and M&S as a methodology is a critical part of the execution of CBT&E.

The shift to a capabilities-based perspective in T&E is not unique to the Department of the Navy. It is happening across Service lines and encompassing all of the Department of Defense (DoD). In 2007, the Deputy Under-Secretary of Defense (Acquisition, Technology, and Logistics) stated, “DT and OT should be integrated and continual to the maximum extent feasible.” The restructuring efforts of the Naval Aviation community lie with the Capabilities Based Test and Evaluation Working Group. Their tasking is to “provide an overarching framework for the development of the guidelines, processes, and procedures for coordination and integration of the Naval Air Systems Command and external organizational capabilities required for the successful execution of CBT&E” (NAVAIRSYSCOM [AIR-5.0], 2011).

The objective of this thesis is to:

- Illustrate the positive effect of incorporating DOE and M&S techniques throughout the entire T&E process
- Quantitatively demonstrate the benefits of CBT&E over SBT&E.

We accomplished these tasks by creating a notional scenario in which a complex joint system defends a Forward Operating Base (FOB). We carried out this scenario using the Situational Awareness for Surveillance and Interdiction Operations (SASIO) simulation model as a proxy for actual live testing. As a secondary objective, we were able to demonstrate the utility of M&S tools for system design and employment. This allowed us to contain all of the myriad T&E

processes across Developmental Test (DT), Operational Test (OT) and Integrated Test (IT) into one succinct, illustrative package, and to present a sample methodology for approaching overall test plan completion.

Capitalizing on previous efforts, we selected DOE as the most effective option for meeting the purposes of T&E. "DOE offers the opportunity to efficiently span major portions of the entire multidimensional test space" (Hutto & Higdon, 2009). We presented the "Plan-Design-Execute-Analyze" conceptual cycle of experimental design, treating this cycle as a roadmap full of guidelines for creating effective test designs. Since no "one-size-fits-all" approach exists in planning defense T&E strategies, this cycle offers a set of mileposts for guiding the DOE process development and effective data analysis in T&E.

We presented an effective design strategy for the DT phase of T&E, illustrating the application of our methodology to determine influential factors in system performance. We proceeded directly to the OT phase, treating results obtained in DT as preferred settings from a design engineer's perspective. We did not initially incorporate integrated testing. Our results indicated system failure in OT resulting from influential factors not considered in the DT phase. We then presented a notional IT phase scenario. This still indicated that the initial test objectives were overly ambitious, but highlighted learning effects and processes gained much earlier (and thus less expensive) in the T&E process.

In summary, this thesis presented the advantage of DOE and M&S in the T&E process, provided a small subset of recommended statistical tools and techniques, and suggested a generalized methodology in the conduct of test plan design. We applied flexible yet powerful statistical techniques in line with the tenets of CBT&E, and can state with confidence that as a methodology, CBT&E will perform no worse, and in most cases substantially better than SBT&E. We presented a brief summary of ongoing work in this field, and suggested possible avenues of further research stemming from this project.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGEMENTS

Earning a master's degree at such a renowned institution as the Naval Postgraduate School represents the feather in my cap as I go forward to face the real world again. This journey would certainly not have been possible without the help and support of many friends and professionals.

I extend a special gratitude and eternal thanks to my advisor, Dr. Rachel Silvestrini, for her continual support and patience throughout this process. This would not have been possible without the expert guidance she has provided. I wish the best of luck to you and your new family, Rachel. I would also like to thank my second reader, CDR Harrison Schramm, for slogging through the muck and mire of my writing and keeping me honest and on track. I look forward to working with you as we continue down the path of military OR!

To Professor Timothy Chung and the rest of your colleagues that developed the SASIO model, thanks ever so much for allowing me to harness the power of your efforts to contribute something meaningful. I would also like to thank the entire faculty at NPS, particularly those in the Operations Research department–it has been a great honor learning from and working with the best!

Without the friends and classmates I have met here in Monterey, this would have been an intolerable journey. To OR cohort 094, it has been my honor working, learning and playing beside you. To the Del Monte Brass, softball, and running crowd, it has been great! Stay on trail, and never lose touch.

Finally, and most importantly, I need to thank my family for your love, support, and most of all, patience. Words cannot describe my gratitude properly. To Kyle and Shelby, Dad is coming home—I'll be there soon. To Tracie, Mitchell, and Mason, I am so happy that we are starting our new life together. It is going to be fantastic!

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Increasingly complex military operating environments have strained the ability of current acquisition processes to field weapon systems that keep pace with technological advances. This has resulted in policy shifts across the Department of Defense (DoD) to mandate a focus on overall system capability to meet a wide range of threats. Traditional Test and Evaluation (T&E) methods narrowly focus on system design to satisfy a particular requirement or performance property, especially in the early developmental phases of design. We refer to this type of testing as Specifications-Based Test and Evaluation (SBT&E). This limits the ability of modern, complex systems to satisfy the capability requirements of the 21st century battlespace, primarily because the emergence of asymmetric threats has driven the Services towards a greater level of interoperability (Defense Science Board Task Force, 2008). The use of SBT&E under these conditions has contributed to the failure of many new systems during the operational test phase.

Service leadership has recognized this trend of increasing failures. In an effort to improve the T&E process, the Department of the Navy (DON) is implementing a focus shift of T&E to a Capabilities-Based Test and Evaluation (CBT&E) process. CBT&E encompasses a broad focus on system design in order to satisfy a particular operational effect that spans the breadth of all phases of T&E. This ensures that the acquisition process delivers operationally effective systems relevant to a wide range of threats and passing the operational test phase.

The CBT&E process will emphasize the design of families of systems to meet the operational commitments of the military communities. These families of systems are a collection of task-oriented systems that pool and integrate their capabilities together to obtain a more complex “meta-system” offering more performance and functionality than the simple sum of individual constituent parts (Popper, 2004). Additionally, CBT&E will increasingly incorporate advanced

scientific and statistical methods, such as Design of Experiments (DOE) and other Modeling and Simulation (M&S) techniques, early and upfront in the design phase. The intelligent application of DOE and M&S as a methodology is a critical part of the execution of CBT&E. The ultimate goal is to field systems that are both fiscally responsible and militarily expedient.

All Services within DoD are transitioning to an employment of CBT&E concepts. Service Operational Test Agency (OTA) Commanders have officially endorsed the idea of this transition, stating that future T&E programs must involve

> …forming a team that must include representation for all testing (Contractor Testing, Government Developmental Testing, Operational Testing), an expert in test design, including DOE, and approval authorities such as DOT&E. (Operational Test Agency Directors & Science Advisor for Operational Test and Evaluation, 2009)

The objective of the research in this thesis is two-fold:

- Illustrate the positive effect of incorporating DOE and M&S techniques throughout the entire T&E process
- Quantitatively demonstrate the benefits of CBT&E over SBT&E.

We satisfy the above objectives by creating a notional scenario in which a complex joint system defends a Forward Operating Base (FOB); we utilized the Situational Awareness for Surveillance and Interdiction Operations (SASIO) simulation model as a proxy for actual testing.

A. BACKGROUND / NAVY INTEREST

In the past, SBT&E was adequate for fielding systems capable of meeting projected threats from potential adversaries. In these operationally diverse yet fiscally constrained times, the advent of complex integrated technologies prevents SBT&E from being effective (Chaudhary, 2000). As outlined by the Defense Science Board (DSB) Task Force in 2008, the challenges of providing new mission systems capable of achieving operational requirements while

meeting time and fiscal constraints confront the Acquisition and T&E communities of all the military Services. For the purposes of this study, we recognize T&E as a required and necessary subset of any Major Defense Acquisition Program (MDAP) in accordance with section 2339 of U.S. Title 10 code. Thus, by default, improvements in the T&E process result in improvements in the entire MDAP process (10USC2399, 2002).

A key difficulty for the T&E community is leveraging resources required to conduct live and rigorous developmental and operational testing. The Director, Operational Test and Evaluation command (DOT&E) has issued guidance for the use of "scientific and statistical methods in developing rigorous, defensible test plans and in evaluating their results" (Gilmore, 2010). The shift to CBT&E stands as the manifestation of Navy and Marine Corps compliance with this directive.

U.S. Navy leadership has identified that SBT&E does not adequately and accurately address the verification and validation of operational effectiveness (OE) and operational suitability (OS) sufficiently early in the development cycle to resolve Critical Operational Issues (COI) (NAVAIRSYSCOM [AIR-5.0], 2011). Furthermore, current T&E methods fail to fully exploit the scope of analytical methods, utilizing such tools as DOE, M&S, and Live, Virtual, and Constructive (LVC) testing (e.g., Hardware-in-the-Loop testing), that could assist in the evolution of a new system acquisition prior to reaching costly and advanced ground and flight activities. Ongoing initiatives in the Naval Aviation Enterprise (NAE) Capability Based Assessment (CBA) Integrated Process (NCIP) recognize the need for robust methodologies to show how multiple complex systems that are collaborative and yet autonomous in nature work together to attain warfighting effects (OPNAV charter [N88], 2009).

The scope of this problem is large, and it not capable of being "solved" in a single document such as this one. In essence, a change from SBT&E to CBT&E is a full-scale "culture-shift" in the T&E community. To manage the transition to CBT&E effectively, in 2011 NAVAIR leadership chartered a collection of acquisition and T&E experts to form the Capabilities Based Test and

Evaluation Working Group (CBTEWG). The mission of CBTEWG is to provide a framework for guidelines, processes and procedures for effective integration and coordination of NAVAIR and external organizational capabilities required for successful execution of CBT&E (NAVAIRSYSCOM [AIR-5.0], 2011). This thesis supports that mission by investigating a small portion of this problem. The work illustrates quantification of gains by applying advanced statistical techniques, through one notional test case study.

An illustrative example of one pending T&E program that will benefit from the improved CBT&E process, including the incorporation of DOE and M&S techniques, is the development of the future Unmanned Carrier Launched Airborne Surveillance and Strike system (UCLASS). Current Navy leadership is highly focused on UCLASS as a this complex System of Systems (SoS) design that will transform the future of carrier-based aviation with an unmanned strike fighter capability that integrates with a multitude of other manned and unmanned weapon systems.

In 2010, the Chief of Naval Operations (CNO) directed UCLASS to be operationally functional by the year 2018; in May 2011, DON awarded a study contract to Boeing to support pre-Milestone A T&E activities in pursuit of operational development (Phantom Works Communications, 2011). However, history has shown that the development of a strike-capable aircraft program requires an average of 17 years from concept to production under the current T&E methodology and approved processes. Successful completion of UCLASS by 2018 is doubtful given the historical precedent. This conundrum is common across all war-fighting communities; solving it will require new and innovative approaches in order to maintain operational relevance in rapidly changing global military environments. This cumbersome T&E process limits our ability to respond to the threat of potential future adversaries, both the conventional state military and non-conventional “fringe group” varieties. The CBTEWG believes that routine incorporation of DOE and M&S techniques during all phases of design and development, particularly in the earliest stages of the process, is one

way of shortening the overall system development process (Standard, Capabilities Based Test & Evaluation: Delivery of Integrated Warfighting Capabilities, 2011).

There are many more situations where the T&E communities across all services could benefit from more integrated testing and in the increased reliance on advanced analytical techniques and simulations early in the design phase. We will highlight the advantages using CBT&E to employ advanced scientific and statistical methods in a rigorous and structured manner in order to identify and manipulate the most important input variables to the process under test. This enables us to illustrate the potential gains of fully integrating modeling, simulation and statistical techniques in all phases of the design and development process.

B. LITERATURE REVIEW

The history of challenges addressing broad capability gaps in the T&E process area is long and varied, with each service within the DoD vying to make their acquisitions processes keep pace with rapidly changing advances in technology. Technological advances have shaped the battlespace in ways that SBT&E methods have failed to predict effectively. In many cases, designers of the legacy systems did not anticipate the need for a capability to adapt to changing threats. For instance, the proliferation of Improvised Explosive Devices (IEDs) in Iraq and Afghanistan was killing a great many service members early on in those conflicts. The designers of the High Mobility Multipurpose Wheeled Vehicle (HMMWV) did not anticipate a need for under-chassis armor, although other fighting vehicles had already employed it. Developers undertook rapid T&E of the Mine Resistant Ambush Protected (MRAP) vehicle to address this change in the adversaries' methods (Atkinson, 2007). The timely fielding of this urgent warfighting requirement stands as a rare success story for the DoD acquisition process (Miller, 2010). Recent history is full of stories of soldiers deploying to the battlefield with equipment technologically 10–15 years or more behind that available through commercial off-the-shelf (COTS) sources. The remainder of

this section tells the story of the long journey towards official recognition of T&E process deficiencies, and highlight specific service efforts to cope with an impending full-scale organizational culture shift.

1. Recognizing T&E Limitations as a DoD Problem

In the March 2000 edition of *Program Manager*, a biannual magazine of the Defense Acquisition University, Capt. Ravi Chaudhary, USAF, published an article highlighting problems with reliability testing in the T&E process. He presents an early argument for incorporating M&S in the integration of Developmental Test & Evaluation (DT&E) and Operational Test & Evaluation (OT&E) specific to reliability considerations, which echoes current CBT&E initiatives. He quotes Dr. George Wauer, Deputy Director for C3I & Strategic Systems at DOT&E as saying, “We can’t afford to wait until OT&E to evaluate system reliability. We need to use system models and testing early enough [before OT&E] to influence the design before changes become too costly” (Chaudhary, 2000).

Paul Davis of the RAND Institute highlighted the tendency of military T&E to focus on individual systems and their requirements individually and without considering interdependencies. His 2002 monograph recommended areas where DoD could change its system of analysis to better support CBT&E. In his words, previous methods were limited to a “bounding-threat method,” where threats at each end of the desired performance range were used as requirements (as represented by one or two point scenarios) which would indirectly lead to the appropriate capabilities. System design was robust to encompass uncertain scenarios requiring flexibility and adaptiveness of capability. Furthermore, this limited design scope led to specific failures by covering an expansive operational envelope that would be better addressed by the growing Family of Systems design approach (Davis, 2002).

Additional commentary on the deficiencies on the DoD T&E process came from Bernard Ziegler and team in their 2005 work “Framework for M&S-Based

System Development and Testing in a Net-Centric Environment" (Ziegler, Fulton, Hammonds, & Nutaro, 2005). They posed the problem as:

> Department of Defense (DoD) acquisition policy requires testing throughout the systems development process to ensure not only technical certification but also mission effectiveness. Complexity within each new system, as well as composition into families of systems and systems of systems, combines with the extensive use of simulation in the design phase to multiply the challenges over traditional interoperability testing methodologies and processes.

This statement captures the essence of the intent of the CBT&E process.

The Office of the Secretary of Defense (OSD) For Acquisition, Technology and Logistics commissioned the DSB Task Force to investigate OSD and Service organizational roles and responsibilities from a T&E perspective. They examined the trend of many programs failing Initial Operational Test & Evaluation (IOT&E) on the basis of being deemed not operationally effective or operationally suitable. While their recommendations were extensive, they present one specific citation relevant to this thesis: "Integrated testing is not a new concept within the Department of Defense, but its importance in recent years has been highlighted, due in part to the growth of asymmetric threats and the adoption of net-centric warfare" (Defense Science Board Task Force, 2008).

Finally, in recognition of the gaps in SBT&E, the Deputy Under-Secretary of Defense (Acquisition, Technology, and Logistics) provided a report to Congress on policies and practices for Test and Evaluation in 2007 (Deputy Under-Secretary of Defense [Acquisition, Technology, and Logistics], 2007). In this report, he stated, "DT and OT should be integrated and continual to the maximum extent feasible," and that "Test and Evaluation should exploit the benefits of appropriate models and simulations." This serves as direct acknowledgement by military leadership of an official change in direction with regard to DoD acquisitions policy.

2. Service Organization Efforts to Comply with the T&E Shift

There are many instances showing the steady progression of an SBT&E shift to the CBT&E process in both open source published literature and within interagency internal traffic. All military services have implemented compliance in shifting T&E focus to CBT&E, corresponding with increasing emphasis on joint interoperability between all governmental agencies. As seen in the OTA MOA, the directors of the five Service OTAs plus DOT&E have endorsed this unique opportunity for rigorous systematic improvement in test processes (Operational Test Agency Directors & Science Advisor for Operational Test and Evaluation, 2009).

The Department of the Navy (DON) has generally recognized the value of M&S (e.g., an early subset of CBT&E) in system design over the course of the last two decades, but implementation has been difficult. In 2002, OPNAVINST 5200.34 stated (Chief of Naval Operations, 2002),

> The Navy adopts and supports the DON M&S vision that modeling and simulation will be a pervasive tool for operational units and will support analysis, training, and acquisition throughout the Department of the Navy.

However, this initial emphasis on M&S generally failed to incorporate the Navy's systems development organizations (e.g., NAVAIR, NAVSEA) and focused on training and deployment simulations at the engagement and campaign levels. In military campaign analysis, system modelers generally incorporate a pyramidal design concept, as seen in Figure 1. The baseline system model (i.e., a single aircraft system) supports an engagement between multiple systems; an engagement model supports a mission or battle involving multiple engagements, and so on up through the theater campaign level. Figure 1 illustrates how DON's emphasis on the goal, a successful campaign or battle, generally neglects the underlying basis of support at the engineering and engagement levels.

The focus of the NCIP effort will ultimately result in a balancing of the modeling pyramid with respect to system design, as illustrated in Figure 2

(Standard, “Naval Aviation Enterprise Capabilities Based Assessment Integrated Process [NCIP],” 2010). T&E processes will deliver a better, more capable product to operators and Navy leadership. Establishing and utilizing a common tool set prior to reaching the Mission/Battle level in this period allows for the implementation of changes early on in the engineering and system design process. This results in a much lower time and monetary penalty.

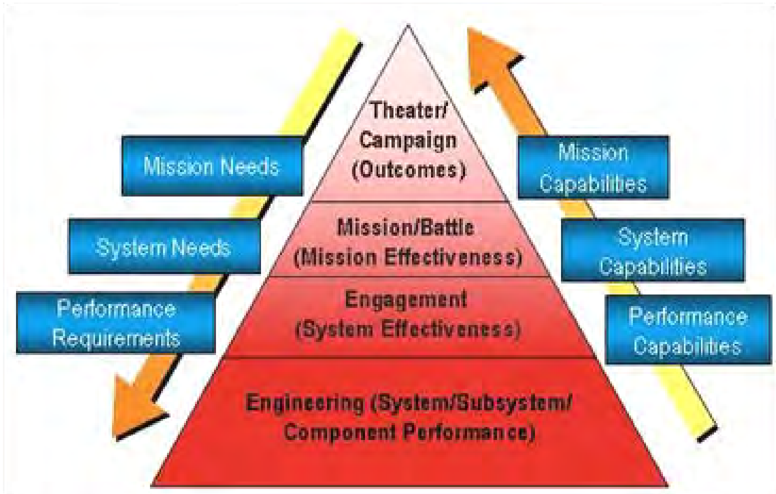


Figure 1. Relationships between the various types of system models and their effect on the overall outcome (From Standard, 2010)

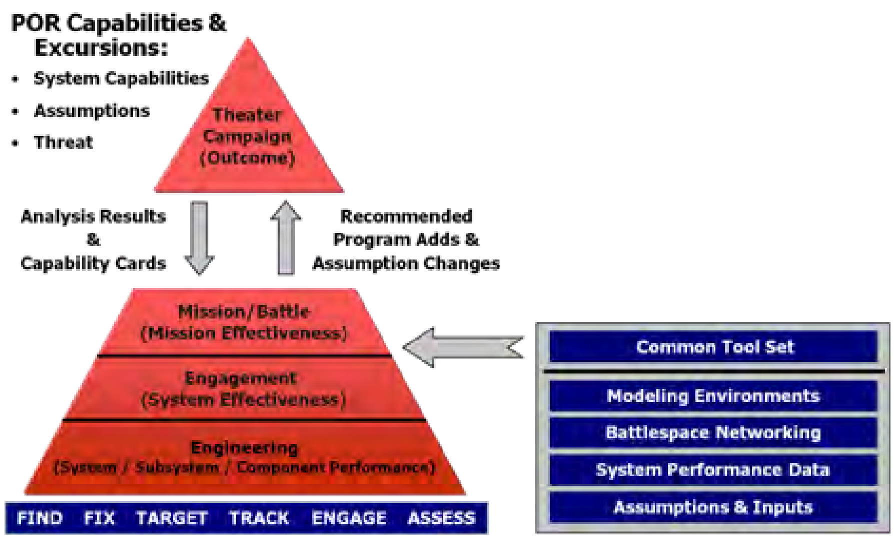


Figure 2. NAVAIR concept of Capabilities Based Test & Evaluation improvements, leveraging M&S in the NCIP program (From Standard, 2010)

NAVAIR has championed the DON's efforts to streamline the T&E process by making efforts to design systems with capabilities in mind rather than specifications. Examples such as the P-8A Poseidon/Broad Area Maritime Surveillance (BAMS) Unmanned Aircraft System (UAS) family of systems and the F-35 Joint Strike Fighter program (in conjunction with the Air Force and Marine Corps) highlight the efforts Naval Aviation has taken. This complies with the NAE vision, the "key to building the force of tomorrow is stabilizing Naval Aviation's investment strategy to acquire the level of warfighting capability and interoperability needed to be successful" (Chief of Naval Air Forces [CNAF], 2010). The capstone of this effort is the formulation of the Capabilities Based Test and Evaluation Working Group (NAVAIRSYSCOM [AIR-5.0], 2011), designed to "provide an overarching framework for the development of the guidelines, processes, and procedures for coordination and integration of the Naval Air Systems Command and external organizational capabilities required for the successful execution of CBT&E."

The U.S. Air Force also addressed the problem in their USAF Early Systems Engineering Guidebook (Assistant Secretary of the Air Force for Acquisition [SAF/AQ], 2009). Specifically, newer warfighting systems are composed of multiple subsystems (e.g., command and control, mission planning, integrated air defense) usually capable of stand-alone operations, that combine to provide an integrated capability. It further states that the integrated SoS capability is the preferred solution over a single weapon system on today's battlefield.

The U.S. Air Force definition of T&E has also responded to this policy shift. Air Force Instruction 10-601 states explicitly:

> The overarching functions of T&E are to determine the operational capabilities and limitations of systems, to reduce risks, and to identify and help resolve deficiencies as early as possible. Integrated T&E combines developmental and operational test objectives to the maximum extent possible and provides assurance

> that systems will satisfy mission requirements in operational environments. (Department of the Air Force, 2010)

This statement illustrates the interservice focus on reorganizing the T&E process.

C. THESIS FOCUS AND ORGANIZATION

Much of the previous literature review focuses on the overarching problem facing the T&E community. Subject Matter Experts (SMEs) have undertaken conceptual studies, working group conferences, and other collaborative efforts on how to improve the process, including parallel work at Commander, Operational Test and Evaluation Force (COMTEVOPFOR). However, we have not found a specific study demonstrating quantifiable gains from the utilization of DOE and M&S techniques. This study addresses that deficiency.

In this thesis, we strive to quantify through an illustrative process T&E enhancements that are possible through the effective utilization of DOE and M&S and other statistical techniques. The Situational Awareness for Surveillance and Interdiction Operations (SASIO) model provides an analysis tool similar to those models utilized by contractors for concept study in the development process. The original purpose of SASIO was to study mission characteristics and performance involving multiple surveillance assets such as Unmanned Aerial Vehicles (UAVs) in conjunction with interdiction assets such as ground-based Quick Reaction Force (QRF) teams. We use SASIO as a surrogate for a *notional* System Under Test (SUT) in a SoS construct in order to address integration and interoperability in the Developmental Test (DT), Operational Test (OT) and Integrated Test (IT) phases of system T&E. Additionally, the nature of SASIO as a simulation allows us to illustrate the use of DOE in a *simulated* environment to predict outcomes in real-world situations.

We organized the remainder of this thesis to best present the use of DOE in M&S and the quantifiable effects of DOE and M&S in the T&E process. In Chapter II, we highlight the specifics of the SASIO model, and present our utilization of SASIO as a proxy for a notional T&E conceptual process. Chapter

III presents an argument for why DOE is the preferred methodology for T&E, and highlights how this thesis illustrates the beneficial effects of advanced analytical techniques in the DT, OT, and IT phases of testing. Chapter IV presents the numerical analysis and results developed to articulate the benefits CBT&E over SBT&E, as well as the development of any tactical or operational insights from the examination of operational teaming. Chapter V provides a summary of this research, contextual comments on relevance to the current problem, as well as recommendations for future research.

II. USING SIMULATION AS A DESIGN TECHNIQUE

In this chapter, we discuss the SASIO simulation model and present its relevance to the operational context. We begin with a general model overview followed by a short discussion of previous work developing and validating its use. We describe the required inputs for running the simulation, and discuss how the model represents the three primary phases of T&E.

We use the SASIO model as a surrogate to represent a live T&E evolution. We treat the outcome of the model as a valid realization for real-world T&E. The initial design concept originally supported either real-time employment strategies (a decision support tool) or robust design strategies (analysis tool) to maximize the employment of surveillance and interceptor assets (Byers, 2010). However, extensive utilization of the model in conjunction with live field experimentation allows us to treat SASIO as a validated and verified combat model for the purposes of analytical exploration. For our analytical presentation, SASIO is a convenient representation of a full-scale operational environment. The concepts we present would work equally well with any similar model or live T&E process.

A. OVERVIEW OF THE SASIO MODEL

The SASIO model is an agent-based stochastic simulation model written by students and professors of the Naval Postgraduate School. SASIO runs using the Java programming language. Researchers originally used the model to simulate a search and interdiction scenario consisting of multiple agents in search of targets, and representative of a notional SUT. It models object motion using Markov transition matrices, object placement through randomized probabilistic mapping, and object location updates through Bayesian updating of the probability map. Reference previous theses by LT Kenneth Byers, USN, (2010) and Maj. Mark Muratore, USMC, (2010) for additional details on the implementation of SASIO.

B. SCENARIO DESCRIPTION

SASIO models an Area of Interest (AOI), such as in Iraq or Afghanistan, in which QRF teams are located at a Forward Operating Base (FOB) and charged with interdicting and capturing hostile targets. Development of the SASIO simulation environment was through ongoing field experiments as part of the USSOCOM-NPS Field Experimentation Cooperative Capabilities Based Experiment 10-3 at Camp Roberts Army Reserve Base. Detection of targets is primarily through Intelligence, Surveillance, and Reconnaissance (ISR) using a UAV Family of Systems (FoS), which we call the Surveyor UAV and the Tracker UAV. The Surveyor UAV performs ISR within the AOI, and has one of three specified search patterns at its disposal. Upon detection of a target, the surveyor sends a report to the QRF, which then proceeds to intercept the target. The Surveyor UAV will either continue to search for additional targets, maintain track through onboard tracking algorithms, or handover tracking responsibilities to the QRF. The QRF can launch an optional handheld Tracker UAV at varying distances from the target location. The model can vary static factors that include search area size, number of neutral and enemy targets, object motion characteristics, and interdictor transit and clear time characteristics. In this case, the Surveyor and Tracker UAVs team with the QRF to locate and capture hostile target entities. In military terms, teaming represents a group of elite soldiers or units, sometimes from different services, working together to achieve a common goal. Figure 3 presents a graphical depiction of the teaming capabilities described above.

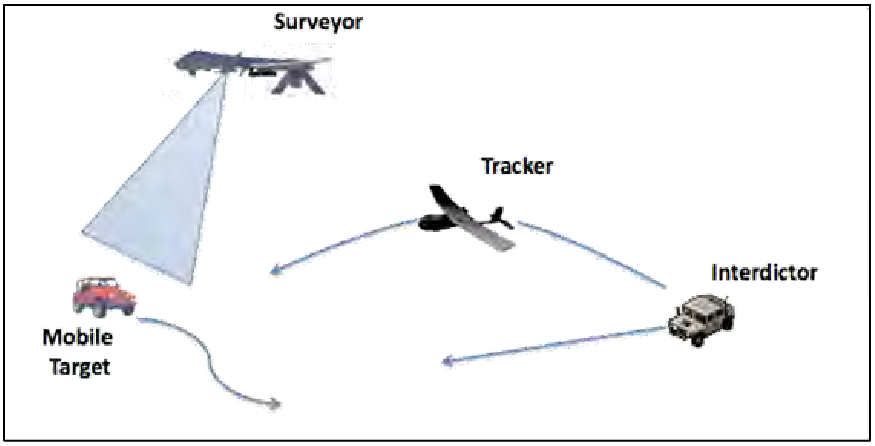


Figure 3. Graphical depiction of SASIO System of Systems teaming capability

The physical realization of the AOI is a portion of Camp Roberts Army Reserve Base, California. The FOB is located at the center of the AOI, and is accessible by three roads or cross-country over the surrounding terrain. Figure 4 depicts the relevant model. Selection of Search Area size within SASIO scales each AOI as an abstraction or extension of the terrain at Camp Roberts. SASIO then treats each AOI as an undirected graph consisting of 1 x 1 km grid squares in which the target is either present or absent (Muratore, 2010).

Figure 4. Tactical Protection of an Installation (From UAV & QRF Barrier Patrol analysis)

C. MEASURES OF EFFECTIVENESS / MEASURES OF PERFORMANCE

We alter many variables (which we call factors) across their ranges of operation (which we call factor levels) throughout the experimentation, which simulates the unpredictability of real-world T&E evolutions. The main objective of the surveillance team and response force throughout testing is to maximize the SUT performance given the available resources. The primary Measure of Effectiveness (MOE) for this SUT is the percentage of hostile targets cleared, with the Measure of Performance (MOP) being the number of hostile targets cleared. We derive the MOE from the MOP as a function of simulation output. Higher numbers of targets cleared indicates more successful system performance. SASIO delivers the metric "Number of Targets Cleared" as a binary response, either one (1) for yes or zero (0) for no. We can convert this metric to a percentage of hostile targets. We use this MOE to gain insights on the best system performance for a given sequence of test scenarios.

D. MODEL INPUTS

The purpose of experimental testing is to determine the specific response of any given process, called the response variable, as a function of various factors and factor levels that encompass the entire factor space. The response variable is synonymous with our MOEs and MOPs. The factors are each associated with levels of operational relevance, such as minimum and maximum speed of advance. Each factor and factor level describes unique characteristics of the entities in the test relevant to the outcome of military operations.

This thesis investigates the factors and their levels as listed in Table 1. Each factor represents a characteristic of the simulation entities and describes a particular value the factor can take during the course of the mission. A detailed description of each factor follows Table 1. Note that due to the nature of the simulation environment, we do not specifically present the units of measure. Where applicable, we relate factor levels to real world units of measure.

Factor	Levels	Type
Surveyor False Positive (γ)	[0, 0.45]	Continuous
Surveyor False Negative (ρ)	[0, 0.45]	Continuous
Search Pattern	Random Walk	Categorical
	Lawnmower	
	Spiral	
Tracker Launch	[1, 5]	Continuous
Tracker Speed	[1, 3]	Continuous
Team Type	Surveyor only	Categorical
	Surveyor with Tracking mode	
	Surveyor w/ Tracker	
Interdictor Transit Time	[15, 1]	Continuous
Interdictor Clear Time	[1, 21]	Continuous
Search Area	[100, 1296, 2500]	Continuous
Number of Objects	[30, 90]	Continuous
Object Motion	Slow Random Walk	Categorical
	Fast Random Walk	

Table 1. List of factors, levels, and ranges

1. Surveyor UAV Factors and Levels

These factors directly represent the characteristic of the Surveyor UAV. Surveyor serves as the core element of the Surveyor/Tracker UAV FoS, and is capable of standalone or integrated operations. We present a description of each factor and their levels of variation.

a. Surveyor UAV Sensor Characteristics

Surveyor False Positive (γ) and Surveyor False Negative (ρ) are continuous factors representing the false positive and false negative detection probabilities of the Surveyor UAV. These factors characterize the imperfect nature of Surveyor's detection capabilities. A false positive detection occurs when the system classifies a target as hostile when it is actually friendly.

Conversely, a false negative detection occurs when the system classifies a target as friendly or neutral when it is in fact hostile. We evaluate Surveyor γ and ρ within the range of [0.0, 0.45].

b. Surveyor UAV Search Pattern

Search Pattern is a three-level categorical variable representing the path motion of Surveyor UAV. Random Walk simulates a flight trajectory of successive random steps and represents a non-systematic random search profile. Lawnmower, more commonly known as Track-line search, represents a systematic search along a pre-planned set of waypoints starting from one corner of a search area and moving across. Spiral, also known as Expanding Square search, represents a systematic search along a pre-planned set of waypoints starting from a center point and radiating outward. For this model, intelligence relating to the target locations does not affect any of the three search patterns.

2. Tracker UAV Factors and Levels

These factors directly represent the characteristics of the Tracker UAV. It serves as the main component of the QRF-launched remote detection capability. The core elements of this system allow the QRF to get visual identification of the target prior to intercept, and thereby release Surveyor (with tracking capability variant) to continue on profile in the detection of additional targets. While Tracker UAV is capable of standalone operation, SASIO does not currently model that functionality.

a. Tracker Launch

SASIO models Tracker UAV using a series of thresholds representing the number of cells away from the target at which the QRF launches it. When combined with Tracker Speed, this influences the response time to the target. This is a continuous factor varied across the range [1, 5] grid squares, representing 1 to 5 kilometers (km) from target in the real world.

b. Tracker Speed

This continuous factor represents the range of velocities that Tracker UAV may fly. The model assumes perfect sensing (i.e., cookie cutter) within its operational envelope. SASIO varies Tracker Speed across the range of [1, 3] time steps per grid square travel, which in the real world is equivalent to 60 kilometers per hour (kph) and 180 kph.

3. Reaction Force and Environmental Factors and Levels

These factors represent the characteristics of the QRF as well as those specific to the operation of the simulation. In an operational environment, QRF teams will have varying levels of proficiency or different means of transit around the AOI. These factors allow the user to control QRF capabilities in order to make the scenario robust to a wide range of operational regimes. Additionally, factors in this category control characteristics specific to the environment in which the T&E evolution might take place.

a. Team Type

This is a three-level categorical factor representing SoS teaming with ground-based QRF assets, and it relates to the employment strategy a commander in the field makes to attain his operational objectives. Level 1 is Surveyor Only, in which the surveyor UAV is the only asset available for operational employment. It cannot track the target, but can only locate and report targets position to the QRF. Level 2 is Surveyor (tracking variant). This is a variation of Level 1, where after locating a target the surveyor UAV maintains target track until relieved by the ground force without a mini UAV. Level 3 is Surveyor w/ Tracker, which represents full teaming capability. Upon target detection, Surveyor UAV maintains target track until relieved by the tracker UAV launched by the QRF.

b. Interdictor Transit Time

This continuous factor is a QRF-related function independent of the Surveyor UAV model. It represents the number of time steps required for the QRF to transit from the FOB to the target. We vary this factor across the range [15, 1] time steps per grid square traveled. The slower speed represents a foot-mounted patrol, while the faster speed represents a vehicle-mounted QRF.

c. Interdictor Clear Time

This is a continuous factor representing a QRF-related function independent of the Surveyor model. We vary it across the range [1, 21] time steps so that it describes the number of time steps required per each interdiction and capture event. The lower clear time represents greater QRF efficiency, and the higher clear time represents a poorly trained unit.

d. Search Area

Search Area is a nominal factor based on AOIs composed of the specified number of 1 km x 1 km grid squares. The AOIs model the scenario environment and directly corresponding to the size of the search area. Larger areas will be more difficult for the SoS to effectively search. Due to symmetry concerns within the model, we limit factor levels to perfect squares (e.g., 100, 1296, 2500).

e. Number of Objects

This is a three-level categorical factor representing the number of entities in a 2:1 neutral to target ratio. For example, value 90 represents 30 hostile and 60 neutral entities. This model varies the factor range on levels [30, 90].

f. Object Motion

Object Motion is a two-level categorical factor representing the self-transition properties of the objects in the simulation. The first level is Slow

Random Walk and the second level is Fast Random Walk. These levels directly represent the characteristics of the hostile and neutral entities in the simulation. Their transitory properties can range from a stationary target to a target that might transition at every time step. The complexity of target detection will vary according to the specified Object Motion setting.

E. PHASES OF TEST & EVALUATION

By treating SASIO as a surrogate for reality, we gain the advantage of using the simulation to manipulate the input factors in a controlled methodology to systematically investigate their effect on the response. We use SASIO as a framework to examine the entire T&E process in small "snapshots." The simulation allows us to split our test runs by controlling the input factors most relevant for each of the DT, IT, and OT phases. Detailed descriptions of DT, OT, and IT follow; however, envision DT as the limited controlled test conducted by a system designer typically in labs and test ranges, OT the full-scale operational employment of the system in a real-world campaign or mission-level context, and IT as the bridge between DT and OT that considers both system design and tactical doctrine. We can run these tests in a progression that best emulates the real-world environment, but also allows us to quantify gains in each phase as well as in aggregate. In Table 2, we expand the factor space outlined in Table 1 to indicate primary or initial test phase of interest.

Factor	Levels	Type	Phase of Testing
Surveyor UAV Factors and Levels			
Surveyor Gamma	[0, 0.45]	Continuous	DT, IT
Surveyor Rho	[0, 0.45]	Continuous	DT, IT
Search Pattern	Random Walk	Categorical	DT, IT
	Lawnmower		
	Spiral		
Tracker UAV Factors and Levels			
Tracker Launch	[1, 5]	Continuous	IT
Tracker Speed	[1, 3]	Continuous	IT
Reaction Force & Environmental Factors and Levels			
Team Type	Surveyor only	Categorical	OT, IT
	Surveyor with Tracking mode		
	Surveyor w/ Tracker		
Interdictor Transit Time	[15, 1]	Continuous	OT, IT
Interdictor Clear Time	[1, 21]	Continuous	OT, IT
Search Area	[100, 1296, 2500]	Continuous	OT, IT
Number of Objects	[30, 90]	Continuous	OT, IT
Object Motion	Slow Random Walk	Categorical	OT, IT
	Fast Random Walk		

Table 2. Expansion of factor space to incorporate the primary test phase of interest

The DT phase is the activity in T&E that focuses on the technological and engineering aspects of a system or piece of equipment. This is where the designer is specifically interested in the product he has been contracted to produce, and is focused on the fine details of its technical performance. In the example we utilize in this thesis, the notional designer is interested in the specific characteristics of the Surveyor UAV, and thus is most concerned with the factors and factor levels of False Positive (γ), False Negative (ρ), and Search Pattern selection (see Table 1). We hold the other factors constant during the DT test at pre-specified levels.

The OT phase represents the culmination of our T&E efforts and includes both controllable and uncontrollable factors in the analysis. Interdictor Clear

Time, Number of Objects and Object Motion are all factors that may vary widely in operational environment. Search Area will also depend on specific employment methods, and Interdictor Transit Type will be a function of Search Area and QRF capability (i.e., foot travel vs. vehicle-borne forces). Team Type will also vary based on the specific employment scenario of the operational unit.

The IT phase is critical to effective CBT&E, as it represents an intermediate collaboration where system designers and operator test organizations learn to share resources and optimize data sharing. The IT phase carries interest for multiple parties, including design teams for each UAV as well as the system operators. This is often the first chance in the overall T&E plan to investigate the effects of teaming assets together as a family of systems. For the purposes of this thesis, we examine the integration of the standalone Surveyor asset with the Tracker UAV carried by the QRF. Therefore, the addition of factors specific to Tracker UAV (Tracker Launch, Tracker Speed) becomes relevant to the analysis. It is important to note that while we treat IT as a separate phase for the purposes of discussion, it actually exists as a continuously updating and repeatable process spanning both DT and OT. For effective CBT&E, OTAs must exploit any opportunity to capture shared T&E events horizontally across organizational lines.

F. RELEVANCE TO THE OPERATIONAL CONTEXT

The T&E evaluation process previously described still focuses on three distinct phases: DT, OT, and IT. We are interested in the use of enhanced analytical techniques early and upfront in the process to enhance the success of operational testing. We utilize SASIO as a surrogate for reality to examine the entirety of the process, and the impact that DOE and M&S techniques have on it.

Authorities in an actual SUT divide the test program into the aforementioned phases. For the purposes of this analysis, we break up the phases to relate to notional program office relationships. The designing program office is concerned with the specific capabilities of the Surveyor UAV; therefore,

factors directly under their control fall in the DT phase. A parallel program office is responsible for the teaming characteristics of the QRF and associated Tracker UAV, including the employment of this asset and its influence on the overall MOE. Finally, the inclusion and examination of all factors, particularly factors external to the SUT design, is extremely relevant to the operational end user.

III. METHODOLOGY

In this chapter, we present our methodology to study the quantitative advantages of incorporating DOE early, upfront, and throughout the design process. We explain the conceptual cycle of experimental design through four primary phases, and then discuss how this cycle is wholly applicable to the T&E process. Primary exploration of this topic is through the presentation of a *notional* SUT and the experimentation we select to investigate its potential. Using SASIO as a proxy for actual testing, we explore the implementation of DOE in the DT, OT, and IT processes.

A. EXPERIMENTAL DESIGN AS THE PREFERRED T&E METHOD

The development, test and evaluation of any particular MDAP is a complex, expensive undertaking; as such, any increases in the efficiency of test design and execution result in considerable savings, both in time and cost considerations. We can measure the cost of test program delays in both the increased expense of the system, as well as in opportunity costs from keeping legacy systems in service longer. There are many approaches to developing and conducting a T&E program. The program consists of many individual phases and sub-phases of design and test events, called experiments, which contribute to the entire process in pursuit of specific system engineering goals. We call the general approach to planning and conducting a series of test event the "strategy of experimentation" (Montgomery, 2009). Methods of selecting the appropriate strategy within each T&E phase include the following:

- The arbitrary selection of factors method (not a scientific process)
- The "best-guess" approach, in which engineers and scientists leverage their experience in the field and subject matter expertise

- The one factor at a time (OFAT) approach, in which a test begins at a baseline point and continues to the end while varying every factor over the range of operations

- The statistical design of experiments approach (also called DOE), which is the process of planning an experiment such that pertinent data is collected and analyzed, resulting in valid and objective conclusions.

DOE offers the best, most effective option for meeting the purpose of T&E: “to mature system designs, manage risks, identify and resolve deficiencies as early as possible, and ensure systems are operationally mission capable” (NAVAIRSYSCOM [AIR-5.0], 2011). Gregory Hutto and James Higdon stated in their 2009 paper:

> Design of Experiments offers an opportunity to improve the way we test – to scientifically justify the number of trials conducted, the arrangement of test conditions, and how to separate the errors in experimental measurement and day-to-day variation from true responses by the system under test. DOE offers the opportunity to efficiently span major portions of the entire multidimensional test space and present those data to the leadership charged with managing the Department of Defense’s $73.2 billion Research, Development, Test, and Evaluation resources in a rigorous, objective manner. (Hutto & Higdon, 2009)

As another testament to the power of DOE in the T&E process, the Service OTA Commanders authored and signed a Memorandum of Agreement (MOA) in 2009 endorsing the utilization of DOE as a common approach in operational T&E endeavors. Specifically, “DOE offers a systematic, rigorous, data-based approach to test and evaluation. DOE is appropriate for serious consideration in *every case* when applied in a testing program” (Operational Test Agency Directors & Science Advisor for Operational Test and Evaluation, 2009). The full text of this MOA is included in Appendix B.

DOE also offers a framework for providing meaningful, scientific answers to the fundamental challenges of any testing evolution. The question of how

many test samples are sufficient to eliminate uncertainty (the number of false positives and false negatives) drives the cost and time required of traditional T&E. Determining which design points (combinations of factor levels for each factor specified in the experiment) to test relevant to DT and OT objectives is a key requirement that DOE can help answer. Planning the execution of the test (in other words, the order in which to perform specific trials), is critical to eliminating bias effects from uncontrollable nuisance variability present in any test environment. Finally, understanding how to draw the appropriate objective conclusions and relate them to specific input/output relationships in order to recommend a course of action is critical to the minimization of time, cost, and risk in the T&E process (Simpson, Hutto, & Sewell, 2011). Figure 5 presents a graphical illustration of the relationship between the inputs (which we call factors), system noise, the process and its resulting outputs (which we call the response variable).

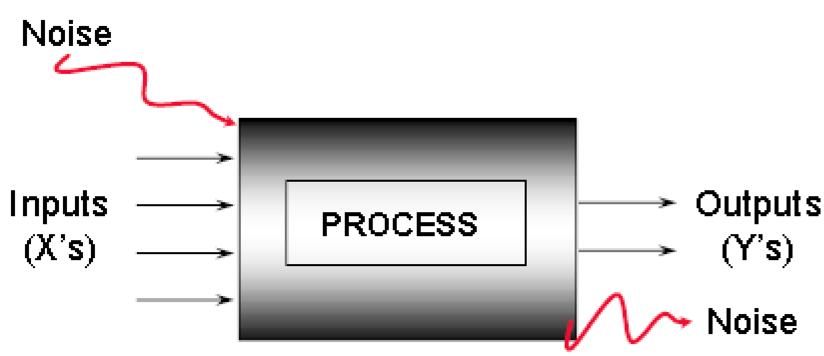


Figure 5. A graphical depiction of the fundamental challenges in experimentation (From Simpson, Hutto & Sewell, 2011)

We can easily relate the SASIO model and its application to T&E to Figure 5. The factors presented in Table 1 serve as our input (X's), which we vary to examine the effect on the output, Percentage of Targets Cleared. It uses Monte Carlo techniques to make SASIO a stochastic process. Figure 6 presents a graphical depiction of the SASIO model.

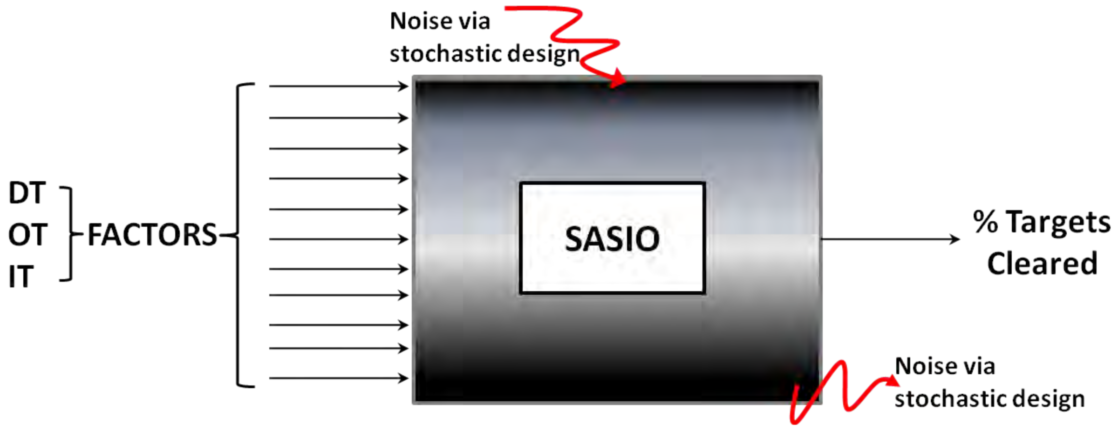


Figure 6. Depiction of the SASIO process

B. EMPLOYING DOE AS A DISCIPLINE TO IMPROVE ALL T&E PHASES

As the history of DOE is rooted in the scientific method, the development of an effective DOE test plan requires the utilization of a scientific approach. Montgomery outlined seven guidelines for designing an experiment (Montgomery, 2009). Steps one through three involve extensive pre-experimental planning. Step four involves the choice of experimental design, taking care to consider the specific objectives of the testing phase and the information required for successful analysis. Step five is the actual execution of the experiment, where errors in procedure could significantly damage the validity of the experiment. Finally, steps six and seven regard the statistical analysis of the data and evaluating practical conclusions for following on courses of action. In the case of T&E, this experimentation exists as a series of iterative processes, with one set of test usually leading to follow-on or sequential experimentation.

In Figure 7, Simpson, Hutto, and Sewell depict the DOE process as a circular cycle of experimentation (Simpson, Hutto, & Sewell, 2011). This cycle follows with Montgomery's guidelines, and provides additional representation for the T&E community. It is not a one-time process, but repeatable across the entire spectrum of T&E phases. In parallel with the Service OTA Commander's policy, the most critical aspect of DOE lies in early and upfront planning encompassing the entire scope of the problem (Operational Test Agency

Directors & Science Advisor for Operational Test and Evaluation, 2009). Furthermore, three of the seven specifically identified uses for DOE from their MOA involve developing a *master plan* for the complete test program, focusing the testing *strategy* across each stage of testing, and iterating *planning* and *testing* correctly to ensure an understanding of the driving factors of system performance (see Appendix B).

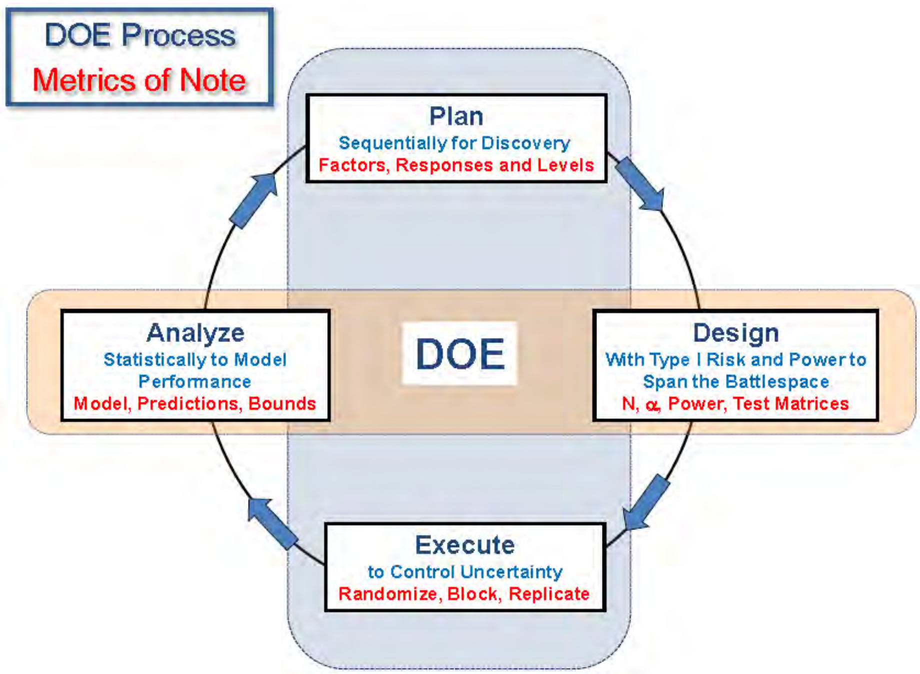


Figure 7. The conceptual cycle of Experimental Design (From presentation, "Embedding DOE in Military Testing: One Organization's Roadmap," Simpson, Hutto & Sewell, 2011)

1. Plan for T&E Success

In the Plan stage of DOE presented in Figure 7, the test authority must involve all stakeholders in the T&E process. This serves to properly scope the objectives of the evaluation across the full spectrum of requirements. One of the most challenging aspects of DOE is identifying the appropriate factors, factor levels and responses to explore.

2. Design for Statistical Confidence

Once the stakeholders have identified the goals of the test program, including relevant input factors and output response variables, the business of selecting the right experimentation scheme must begin. The Design stage should consider aspects of T&E important to the Program Manager, the system developer, and the end user (operator). In DOE, design consists of the selection of trials (also known as design points) that make up the experiment, where each trial is a selection of a factor level corresponding to each input factor of interest.

Selection of the proper design is not an easy matter, nor is there a checklist that applies universally to every situation. DOE is a balance of budget, time, efficiency, and adequacy of design to cover the factor space. Such items as number of observations (sample size), number of repeated experiments (replicates), total test cost, randomization of observations, and suitable run order are critical to the successful implementation of the Test and Evaluation Master Plan (TEMP). The design also needs to incorporate the idea of statistical confidence and power across the battlespace. Statistical confidence contains the desirable effect of minimizing false positives leading to unnecessary system overload, while statistical power involves the minimization of false negatives, which could be extremely detrimental in a battlefield environment (Hutto & Higdon, 2009).

A wide variety of standard techniques and commonly used designs exist that can be tailored for T&E situations. There is an extensive body of work comparing the relative merits of various designs. Specific designs that work well for T&E processes include factorial designs to study the combined effect of factors on a response, fractional factorial designs (when the number of required experiments grows beyond acceptable resource levels), and optimal designs for the attainment of specific goals. Refer to texts by D.C. Montgomery (*Design and Analysis of Experiments*, 2009) and software packages such as JMP 9 (SAS Institute Inc., 2011) to find detailed discussions of these types of designs. These

designs are useful to meet the goals of T&E experiments and we can execute them easily in simulation as well as live experimentation.

3. Execute for Test Plan Success

The Execute stage is the actual performance of the selected experiments in the simulation or during live experimentation. Operators should take care in this stage not to undermine the careful planning executed in stage 1 through careless errors in performing the experiment. In this phase, it is important to control the effects of uncertainty through the proper application of randomization, replication and statistical blocking techniques. Randomization and replication refer to the order and number of times in which we test specific design points within the experiment. They help prevent unknown effects from influencing the results while aiding in the estimation of the variability. Statistical blocking is the practice of arranging experimental units in groups (blocks) that are similar to one another. It serves to reduce unintended sources of variability so that we may confidently state that the variability in the response factor is due to our selection of inputs rather than an unexpected combination of effects.

4. Analyze for Meaningful Decision-Making

In referring to the Analyze phase, we mean the mathematical application of statistical methods to the data collected in the Execute phase of DOE (evaluation of the conduct of the physical experiment must be done, but separate and independent of the analyze phase). In this phase, test authorities apply objective statistical methods to provide analytical rigor while avoiding Service, Community or personal bias in the presentation of results. These results allow for a measure of likely error or level of confidence in our conclusion important for the decision-making process. Utilization of results in the design of follow-on experiments allows us to “accumulate evidence that the system performs across its operational envelope” and to formulate “meaningful integrated testing” (Operational Test Agency Directors & Science Advisor for Operational Test and Evaluation, 2009).

C. STATISTICAL METHODS OF ANALYSIS

We use well-established basic and advanced statistical methods in the Analyze phase of DOE to ensure objectivity in the results and conclusions of the test. This prevents subjectivity and human preference from unfairly biasing the results. These statistical methods do not prove or disprove that a factor has a particular effect. They do, however, provide a measure of the likely error in a given conclusion or enable us to attach a level of confidence to a statement. We provide a brief overview of several of these important techniques in the following section. For a full discussion of these statistical techniques, refer to texts by noted authors Douglas Montgomery (*Design and Analysis of Experiments*, 2009), Jay L. Devore (*Probability and Statistics for Engineering and the Sciences*, 2009) and others for detailed explanations.

1. Analysis of Variance

Analysis of Variance (ANOVA) is a statistical procedure that is very effective in analyzing highly structured experimental data. We use ANOVA to provide a measure of the relative variability between sets of models fit to a particular set of data. We use ANOVA to describe the classic linear model represented as a decomposition of data into a grand mean, main effects, possible interactions, and an error term. This decomposition allows us to estimate variation resulting from individual components of the model. We may then compare the observed data to a reference distribution (in this case the *F*-distribution) to compare our model components against the hypothesis that any source of variation in the model is zero.

We also apply ANOVA techniques to multivariate linear regression models and generalized linear models (like logistic regression) to compare regressions with large numbers of predictors. Finally, we use ANOVA to compare multiple models to determine if a simpler one is sufficient to explain the variation (Gelman, 2005).

2. Multivariate Linear Regression

We use statistical methods to determine what factors in an experiment have a significant effect on the response variable. We use multivariate linear regression to characterize the relationship between these variables, called regressor variables, with a mathematical model fit to a set of sample data. We then use this model to approximate the response for any given set of input data. The standard multivariate linear regression is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{1}$$

where *k* is the number of regressor variables.

In Equation (1), y represents the response variable and the x_i's represent the regressor variables for each factor in the test design. The parameters (β_j), called partial regression coefficients, represent the expected change in the response *y* per unit change in x_j when we hold all of the other regressor variables at constant value. The statistical error, ε, represents the difference between the observed value of the experiment and the predicted value y. Additionally, Equation (1) specifies a model containing only the main effects from each factor plus the aggregated error term. We can also expand it to incorporate multi-factor interactions or quadratic terms when necessary.

We use the standard multivariate linear model to test the following statistical hypothesis, as depicted in Equation (2):

$$\begin{aligned} &\mathrm{H}_0 : \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_k = 0, \text{ for } k \text{ number of regressors} \\ &\mathrm{H}_1 : \beta_j \neq 0 \text{ for at least one } j \end{aligned} \tag{2}$$

In general, a regression model that is linear in the parameters (the β's) is a linear model regardless of the shape of the response surface generated (Montgomery, 2009).

In most situations, multiple regression models are easier to work with when presented in matrix notation. In this method, the multivariate linear model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{4}$$

This method of presentation allows for a very compact display of the data and results gathered from the T&E experiment we are conducting. Note that Equations (3) and (4) include only the main effects from each factor. We can expand the matrices to include additional terms in the model (i.e., two-factor effects, quadratic effects) by adding columns to the $\mathbf{X}$ matrix for factors and adding rows to the $\boldsymbol{\beta}$ matrix for partial regression coefficients (Montgomery, Peck, & Vining, 2006).

3. Logistic Regression

We use the statistical technique called logistic regression to predict the probability of an event. Logistic regression is a form of the generalized linear model (GLM). The GLM takes a non-linear response and generalizes it to the standard linear regression by relating the response variable to a link function. The link function provides the relationship between the linear predictor and the mean of the distribution function. This, in turn, allows the magnitude of the variance of each observation to exist as a function of its predicted value (Montgomery, Peck, & Vining, 2006).

For this thesis, we are interested in the probability (or percentage) of hostile targets cleared from an Area of Interest (AOI). The SASIO model represents the event of clearing a target as a Bernoulli random variable that can

take on a value of zero or one. The target is cleared (value of 1) or it is not cleared (value of 0). Since the response variable is binary, then the shape of the response function is non-linear. In logistic regression, this non-linear response function is very popular and takes the name "logit." We present the form of the logit function in Equation (5):

$$\text{logit} = \frac{1}{1+e^{(-\mathbf{X'\beta})}} \tag{5}$$

We then "linearize" the logit response function via a technique called Log-Odds, which is a transformation back to a linear form compatible with standard multivariate linear regression (Montgomery, Peck, & Vining, 2006). This transformation occurs through use of the logit link function (**y***), which takes the form

$$\mathbf{y}^* = \ln\left(\frac{\mu}{1-\mu}\right) = \mathbf{X\beta} + \boldsymbol{\varepsilon} \tag{6}$$

where μ generically represents the parameter we wish to transform. Notice that Equation (6) is equivalent to the resultant of Equation (3). In our case, SASIO determines the mean number of targets cleared, which we transform to the mean percentage of targets cleared (μ) by dividing by the number of targets for that design point. We are then able to perform the standard multivariate linear regression for analysis.

D. DOE AS IT APPLIES TO TEST AND EVALUATION (T&E)

Thus far, we have discussed the DOE methodology and statistical analysis techniques with some specificity, but in very general terms as they apply to T&E. Of particular interest to this thesis, however, is how incorporating DOE and M&S techniques throughout the CBT&E process positively improves a T&E program from previously utilized SBT&E techniques. We present how the continual implementation of Plan-Design-Execute-Analyze methodology may enhance overall process results.

1. DOE in the Developmental Testing (DT) Phase

The first phase of any T&E program is the DT phase. Simply defined, DT is T&E "conducted to measure progress, usually of component/subsystems, and the proofing of manufacturing processes and controls and to assist the engineering design and development process and verify attainment of technical performance specifications and objectives" (Development Test and Evaluation [DT&E], n.d.). DT focuses primarily on the private and governmental contractor level requirements of system design. This phase is critical in evaluating the specific SUT for risk level information, risk mitigation techniques, the feasibility of technical performance parameter attainment, and data collection for model and simulation validation for use in later phases of testing.

The application of DOE as a specific test methodology is relatively straightforward in the DT phase. Complexity in DT test design is directly proportional to the number of input factors involved in the experiment. In complex systems with many input factors and multiple response variables, the intelligent application of DOE or similar techniques is critical to avoid costly redesign and rework. DT lays the foundation for follow-on testing for suitability in the operational environment.

The most critical aspect of DT is the ability of the designer to control the test environment and the factor space. The ranges and/or levels of each factor, including multi-dimensional combinations and interaction of factors and process parameters, define factor space. The SUT developer controls which factors affect the SUT so he may accurately measure the desired response variable with limited statistical bias. His objective is to identify the preferred settings of input factors, such as sensor performance characteristics, than allow for system performance sufficient to satisfy test requirements. Unfortunately, the controlled environment does not always accurately represent the full operational envelope of the SUT.

Due to the nature of the systems engineering process, DT tends to focus on attaining specific technical performance parameters. Such technical specifications might include the precision of a radar system (i.e., detection of a one square meter cross-section at 200 km) or the fidelity of an observation camera. As the core concept behind SBT&E, this design technique is convenient and unambiguous in assessing the functionality of the system, and has worked sufficiently for many years. However, in today's more technical and interoperable world, this method may not directly translate to the focus on overall operational performance necessary in CBT&E. In other words, the true objective of CBT&E is to field a system (or system of systems) that can contribute to mission accomplishment at the appropriate level, across a broad range of performance levels. In the radar example, it may be sufficient to specify that the system must be able to detect and report an inbound target in sufficient time such that a friendly unit can engage it before the target is able to use its weapons against us. We will explore this idea with the Surveyor/Tracker FoS.

DT needs to use sound experimental methodology to develop target factor levels for follow-on experimentation. We will use DOE in this fashion to illustrate gains attainable early in the design process. The system capability for our illustration is the following: "The Surveyor/Tracker UAV FoS should be able to reliably detect and capture, in conjunction with a well-trained QRF, at least 85% of enemy targets in an environment suitably representative of operational conditions." In this thesis, our DT phase focuses exclusively on the capabilities of the Surveyor UAV as described in Chapter II.

2. DOE in the Operational Testing (OT) Phase

The final phase of T&E for a military weapons system is the OT phase (typically referred to as OT&E). A good working definition of OT is as follows:

> That T&E conducted to estimate a system's military utility, operational effectiveness, and operational suitability, as well as the need for any modifications. It is accomplished by operational and support personnel of the types and qualifications expected to use

> and maintain the system when deployed, and is conducted in as realistic an operational environment as possible. (Operational Test and Evaluation [OT&E], n.d.).

The OT phase is considerably more complex than the DT phase. The number of input factors, as well as the number of response variables, usually increases substantially from the DT phase. The inclusion of uncontrollable factors, appearing in our mathematical model as *noise*, such as environment, varying operating envelopes, multi-system interoperability requirements and relatively untrained operators, may have a significant effect on the ability of the SUT to accomplish the desired operational capability. Increased and well-applied M&S helps alleviate some of these issues by helping to mitigate known resource limitations and by providing technical and programmatic risk reduction.

Due to the above considerations, successful accomplishment rates (i.e. pass/fail in OT) often drop significantly. Systems and input factor levels that worked exceedingly well in DT may not work well at all once exposed to a more expansive factor design space.

We illustrate this phenomenon by using our model to increase the size of the process design space (cover more factor levels with a higher number of experimental trials) for the OT phase. Initially, we treat the factors examined in DT as fixed from predetermined levels, which represent the optimal specification settings of the Surveyor UAV. We also illustrate the application of DOE methodologies to an expanded range of input variables. For example, we expand a controlled AOI size from the DT phase to meet the needs of the operational commander in the field and demonstrate the impact on system performance.

It is easy to explore a relatively small series of experiments using full-factorial designs. In our DT scenario, testing with three primary input factors only requires a minimum of 12 design trials to fully represent the design space (a high and low level for each sensor probability γ and ρ tested at each of the three search pattern settings). However, as the complexity of testing increases (as in

OT) this is not possible due to time, budget and resource constraints. The cost to properly test and evaluation complex systems can run into many millions of dollars and many hundreds of person-hours.

One advantage to this study, however, is our ability to establish a baseline set of results by actually investigating the entire factor space. We present these results as an academic comparison of readily achievable results through smart application of DOE versus the full set of 4,608 design points (which is the number of trials required by a full-factorial design in all factors).

The complexity of highly interoperable systems also highlights challenges within the highly competitive T&E community. Historically, Service components have developed mission systems unique to the requirements of their mission requirements. For example, the Navy needed heavy fighter aircraft that were highly maneuverable but could still withstand the stresses of aircraft carrier launch and recovery. Thus, the F-14 Tomcat was specific to the USN. The Air Force F-15 Eagle heavy fighter aircraft is also highly maneuverable and has many similarities to the F-14, but did not need to land on a sea-based platform. Despite the extreme system similarities, the programs the programs developed completely independent of each other. Many people would argue that significant cost and resource savings by developing systems for joint use across the Services. Certain historical, structural, organizational and even legal barriers prevent the free-flow of data amongst the OTAs. Management professionals commonly refer to these barriers as “stovepipes” within an organization, which characteristically restrict the flow of information to up and down lines of control and inhibits cross-organizational information sharing.

This type of organizational structure applies in many aspects within the T&E community. Data sharing is limited between DT and OT organization by precedence, lack of communication, or burdensome bureaucratic processes. The implementation of the CBT&E process within Navy OTA channels is one effort to reduce stove piping within the organization and more effectively share

data and resources. Sebolka, Grow and Tye present a good discussion of this organizational culture in their 2008 International Test and Evaluation Association article (2008).

3. DOE in the Integrated Testing (IT) Phase

In recognition of the excessive cost and timelines associated with T&E, the T&E community leaders mandated the use of integrated T&E in December 2007. The Office of the Secretary of Defense issued the following definition of integrate testing (IT) in April 2008:

> Integrated testing is the collaborative planning and collaborative execution of test phases and events to provide shared data in support of independent analysis, evaluation, and reporting by all stakeholders, particularly the developmental (both contractor and government) and operational test and evaluation communities.

Organizational T&E authorities intend for increased utilization of IT to transcend some of the stovepipe mentality that does exist in order to capitalize on cost, time and risk savings within the community. A single test event for OT and DT has the potential to answer both DT and OT questions efficiently in terms of the time and resources required when properly applied.

To be very clear, by the definition IT is *not* a separate test phase or new type of test. It is a *process change* meant to result in robust data sharing amongst test organizations. This process change makes IT a major component in the entire CBT&E process, as one cannot design a complex SUT to accomplish a particular capability unless it functions well with the other components of the SoS. For example, data stemming from integrated testing might allow the contractor to improve his basic design (e.g., Surveyor UAV), the DT evaluators to assess risk, and OT authorities to conduct initial operational assessments.

For the purposes of this study, however, we treat IT as an intermediate, separate phase that exists between initial DT and final OT. The IT phase exists as an iterative process. The Army Test and Evaluation Command (ATEC)

relates this to the “model-test-model” approach, which is used throughout the acquisition life cycle to effectively focus T&E resources on critical test issues (Streilein, 2009). Not only does Dr. Streilein promote the idea behind IT, he also emphasizes the critical relationship between live testing and M&S within the T&E process.

Within this thesis, we use the idea of IT as an individual phase to show the DOE and analysis methods that an experimenter must apply on all levels of T&E to enable a true CBT&E approach. We combine the use of both DT and OT factors within our design space to show the power of DOE methods. We use the results from DT as starting points for sequential experimentation, just as a tester should do in an actual test evolution. Additionally, we present a notional SoS for testing, including specific operational aspects (such as teaming, UAV integration, and environmental flexibility), that represents the entirety of the notional T&E process.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ANALYSIS AND RESULTS

In previous sections, we explained the history of T&E and demonstrated theoretical reasons why a DOE methodology may be superior to other approaches. We are now ready to demonstrate the advantages with a concrete example. In our notional scenario, we used percentage of targets captured as our MOE, which serves as a "benchmark" to compare the military efficiency of various designs. We used DOE to determine which significant factors affect the response variable. Additionally we aimed to find desirable factor levels for Surveyor False Positive (γ) and False Negative (ρ) probabilities that present the best chance to meet operational effectiveness and suitability criteria in the OT evaluation.

In this chapter, we present selection of the design and the analysis for the DT, OT and IT phases of our example T&E scenario. For each circumstance of DT, OT, and IT phasing, we present the Plan, Design, Execute, and Analyze methodology.

A. CONSIDERATIONS FOR PROPER IMPLEMENTATION OF DOE

In Chapter III, we introduced the concepts of sample size, randomization, and replication as critical elements of experimental design; this remains true in the implementation of our design for the T&E process. Sample size refers to the number of observations (or design points) used to evaluate the MOE, and directly contributes to the accuracy of our analysis as well as the overall cost of the event. Randomization prevents the inadvertent introduction of error into the process from nuisance or unaccounted for variables, or uncontrollable environmental factors, such as weather, terrain or ground clutter. While these uncontrollable factors are not present in the sanitary environment of a computer simulation, we retain randomization of design points so that our simulation matches our proposed live T&E methodology.

An important aspect of experimentation is to identify the sources of error and uncertainty within the process. In a stochastic simulation model, such as SASIO, individual runs of any given design point might result in an unstable variation between results. To counter this potentially negative effect, we conducted 60 replications of each design point within SASIO. Previous research has shown that 60 trials is the number of replications that stabilizes the variance inherent to the model (Muratore, 2010).

One additional characteristic of DOE not previously mentioned but extremely important to attaining valid results in T&E is the concept of statistical blocking. In certain cases, such as in flight-testing employing several different pilots or in missile testing where the test articles may come from different manufacture lots, factors outside the design may introduce unwanted variability. This nuisance variance may prove detrimental to the test results. We can control this through blocking, which is the grouping of experimental units into blocks that are similar to one another. We then confine our comparisons to those within the blocks, thus attaining greater precision by eliminating the difference between the blocks (Montgomery, 2009). Because we do not have uncontrolled variables in the computer simulation, blocking is not required; however, we once again mention this design concept to maintain relevance to live T&E events.

Finally, every T&E event has a specific response variable (or response variables) of interest to the test authorities. This objective measurement, or MOE, is central to the selection of an appropriate design for each event. To reiterate the purpose of our T&E methodology, “The Surveyor/Tracker UAV FoS should be able to reliably detect and capture, in conjunction with a well-trained QRF, at least 85% of enemy targets in an environment suitably representative of operational conditions.”

B. EXPERIMENTAL DESIGNS AND RESULTS BY PHASE

In this section, we present the designs and experimental results by phase in order to illustrate the effectiveness of DOE and M&S in the T&E process. We designed our notional scenario to draw out certain aspects of the different T&E approaches to make specific comparisons; real-life experiments almost certainly have more variables. This section serves to highlight an appropriate application of the DOE methodology. We focus on the selection of design and analysis of results. We identified factors and levels in Chapter II, Table 1.

1. Developmental Test (DT) Phase

The first part of the methodology deals with the *notional* DT phase, and how three primary factors affect the percentage of targets cleared. The notional program office responsible for designing the Surveyor UAV portion of our Family of Systems controls the manipulation of these factors within the laboratory or on the test bench. In traditional T&E methods (i.e., SBT&E), attainment of well-defined system specifications and key performance parameters in a controlled environment is the overall goal of the DT phase. In our model, we focus on the attainment of a defined capability, seeking the design parameters (factors) that enable that objective.

a. Planning and Design Considerations in DT

In accordance with the conceptual cycle of DOE (Plan-Design-Execute-Analyze) presented in Figure 7, the first step is to make a comprehensive test plan. This involves obtaining input from every stakeholder in the process to determine specific objectives, factors and response variables important to the test. The specific objective for this phase of T&E is to determine the preferred sensor characteristics (γ and ρ) and search pattern to employ in order to capture at least 85% of the hostile targets that ingress the AOI. We treat capture percentage as a *capability* required by the field commander to enable mission accomplishment, which is to protect the FOB from hostile takeover.

Considering the objectives and inputs determined in the planning phase, the second phase is to design the experiment that will best attain those objectives with minimum cost and time. Since the number of factors is relatively low, we selected a $2^2 \times 3$ full factorial design augmented with two center points on each face, resulting in 18 design points. We illustrate this design graphically in Figure 8. Researchers often use this type of design as a screening experiment, where the goal is to determine preliminary information about significant factor effects. In particular, this design goal serves as a good choice for DT. Testing at the endpoints for each factor level allows for a complete examination of the factor space. The augmentation of the design with center points allows the experimenter to test for any quadratic effects in the model, as well as independently estimate the true error within the design (Montgomery, 2009).

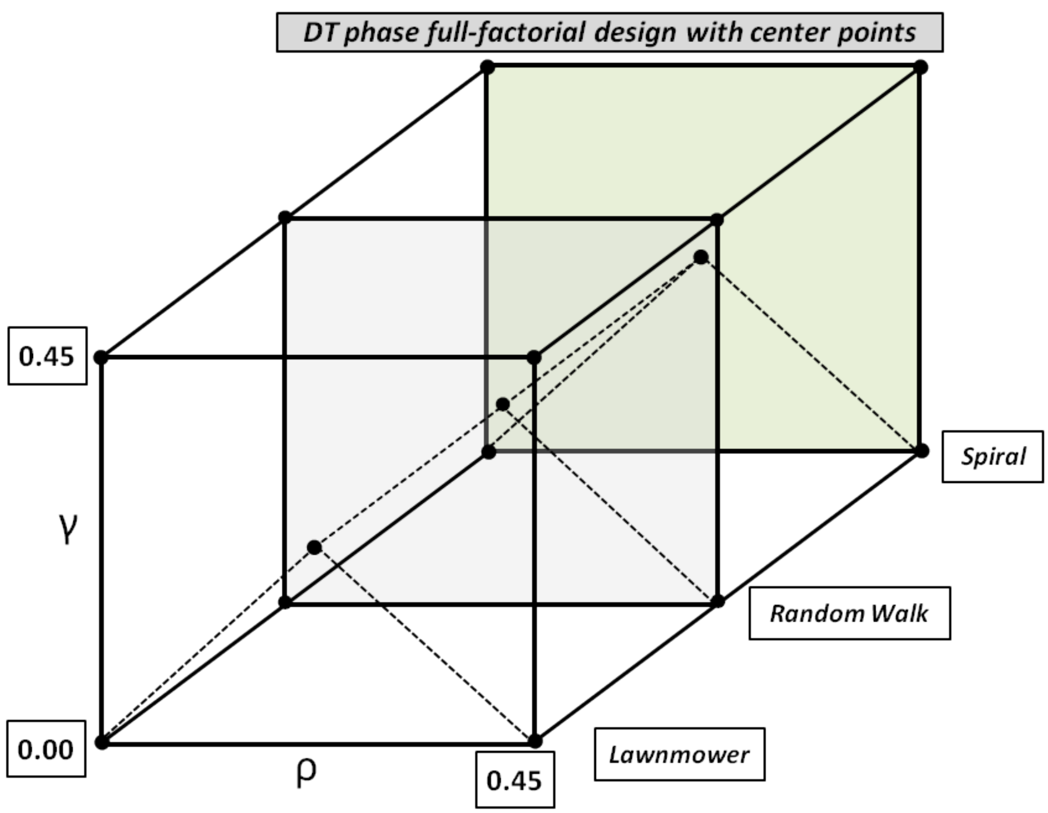


Figure 8. Graphical presentation of the DT phase experimental design

With a design in place, we (the experimenter) execute the design in a controlled environment. In our case, since we are using a simulation we needed to include all factors in the model to run SASIO. To control this environment, we hold the factors not critical to the test constant at pre-specified levels. Table 3 presents our held-constant factor settings for the DT design. Note, however, that if necessary to the attainment of the design goal for any particular test regime, we may vary these factors in each simulation run as part of the overall design.

Held Constant Factors	Setting
Team Type	Surveyor / Tracking
Tracker Launch	3
Interdictor Transit Time	1
Tracker Speed	3
Search Area	100
Clear Time	1
Number of Objects	30
Object Motion	SlowRW

Table 3. DT Phase Held Constant Factors and Factor Levels

Table 4 presents the experimental design in the test factors of interest (grouped by search pattern for clarity). There is a one-to-one correlation between this table and Figure 8. Notice that the "Design Point" indicates the random sequence in which we executed that particular trial. This randomness is necessary in live test events for the reasons stated above and is included here for completeness. For each search pattern, we tested γ and ρ at all combinations of their high and low factor levels. Additionally, we ran two observations at the midpoint of each test face; Figure 8 graphically depicts the test faces.

Design Point	False Pos. Prob (γ)	False Neg. Prob (ρ)	Search Pattern
1	0.450	0.000	Random Walk
6	0.000	0.000	
8	0.225	0.225	
10	0.450	0.450	
11	0.225	0.225	
17	0.000	0.450	

Design Point	False Pos. Prob (γ)	False Neg. Prob (ρ)	Search Pattern
2	0.000	0.450	Spiral
4	0.450	0.450	
5	0.000	0.000	
13	0.225	0.225	
16	0.450	0.000	
18	0.225	0.225	

Design Point	False Pos. Prob (γ)	False Neg. Prob (ρ)	Search Pattern
3	0.450	0.450	Lawnmower
7	0.225	0.225	
9	0.000	0.450	
12	0.450	0.000	
14	0.000	0.000	
15	0.225	0.225	

Table 4. DT Phase Experimental Design factors, grouped by Search Pattern

b. Execution and Analysis of DT Results

Once the ‘execute’ phase is complete, the statistical analysis phase begins. We used the JMP Pro 9.0 software package (from SAS Institute, Inc.) to create the experimental designs and analyze of the data. As developed in Chapter II, we used logistic regression with a logit link function to map our Bernoulli response onto a linear regression. The operational importance of the Logit is not obvious to most customers; therefore, we transform the results back to percentages for discussion and reporting. In Equation (7), we apply the logit link function to evaluate an 85% target clearance rate.

$$\mathbf{y}^* = \text{logit}(85\% \text{ targets cleared}) = \ln\left(\frac{0.85}{1-0.85}\right) = 1.7346 \tag{7}$$

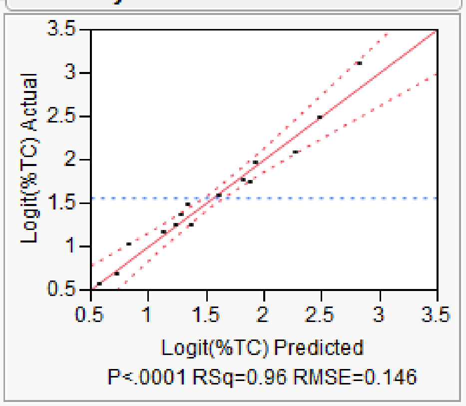
As you can see, a response of at least 1.7346 from our predictive logit model corresponds to an 85% target clearance rate. When necessary, we transform logit values back to percentage form using Equation (8), where $\mathbf{y}^*$ represents our predicted response (logit [percent targets cleared]).

$$\text{Percent (targets cleared)} = \frac{e^{y^*}}{1+e^{y^*}} \tag{8}$$

Keep in mind that any simulated average that achieves this value is a point estimate. Statistical methods provide a confidence interval about this estimate to account for statistical error. For simplicity of discussion, we treat the point estimate as valid statistical criteria. In an actual SUT with a strict requirement for system performance and reliability, we could force the design to ensure that the lower confidence level limit satisfies the 85% MOE. However, for ease of presentation we use the average.

In our analysis, our best logistic regression model showed that three main effects and one interaction were significant. Figure 9 presents a summary of our results. We look at R^2-adj., which is the coefficient of multiple determination adjusted for the number of factors in the model, as a metric for comparing competing regression models. It states that our model is sufficient to explain approximately 94.3% of the variability.

Actual by Predicted Plot



Summary of Fit

RSquare	0.963065
RSquare Adj	0.942919
Root Mean Square Error	0.14604
Mean of Response	1.555303
Observations (or Sum Wgts)	18

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	6.1172515	1.01954	47.8035
Error	11	0.2346054	0.02133	**Prob > F**
C. Total	17	6.3518568		<.0001*

Figure 9. Linear Regression model of LOGIT transformation of Percent Targets Cleared

Figure 10 lists the parameter estimates of our fitted model. The last column includes an estimation of significance to the model of each factor effect called *p*-value. *P*-value is a statistical metric used to determine the relative importance of each factor in the model. Specifically, if our null hypothesis were true (that our regression coefficients equal zero; see Chapter III), *p*-value is the probability of observing a test statistic at least as extreme as the one we observed. A *p*-value lower than a specified significance value (α) indicates that particular term is influential to the model or process under test. In this model, we see that the greatest negative effect comes from a high false negative probability rate. This is consistent with our operational intuition, because classifying a hostile as friendly potentially poses a great danger to the force. Of particular operational employment consideration is the negative significance of the Random Walk search pattern. As we examine later, this turns out to be the least effective of the three available search patterns.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.3995699	0.068844	34.86	<.0001*
Search Pattern[Lawnmower]	0.0848292	0.07697	1.10	0.2939
Search Pattern[Random Walk]	-0.512892	0.07697	-6.66	<.0001*
False Pos. Prob	-1.226786	0.18737	-6.55	<.0001*
False Neg. Prob	-2.525514	0.18737	-13.48	<.0001*
Search Pattern[Lawnmower]*False Neg. Prob	-0.147804	0.264981	-0.56	0.5882
Search Pattern[Random Walk]*False Neg. Prob	0.8465956	0.264981	3.19	0.0085*

Figure 10. Parameter Estimates of LOGIT Transformation of Percent Targets Cleared

The planned objectives of the DT phase were to ensure adequate Surveyor performance, and to determine the best settings to attain at least 85% target clearance. We transformed logit values to percentage form for further evaluation. Figure 11 presents contour plots for predicted percentage of targets cleared (based on our fitted model) for each search pattern, contrasting false negative probability (ρ) against false positive probability (γ). The contours

represent the combination of γ and ρ that the designer could choose from to meet the desired system capabilities. Given the conditions of the DT, he determines the optimal settings for Search Pattern, γ and ρ to attain at least an 85% predicted number of targets cleared. For instance, in the Lawnmower Search Pattern plot, the 0.850 contour line runs from a point at a (γ, ρ) coordinate of (0.0, 0.23) to a point at (0.45, 0.06). Any point on or below this line represents a (γ, ρ) combination where the average percentage of targets cleared satisfies the 85% MOE. We clearly see that Random Walk performs the worst in this environment.

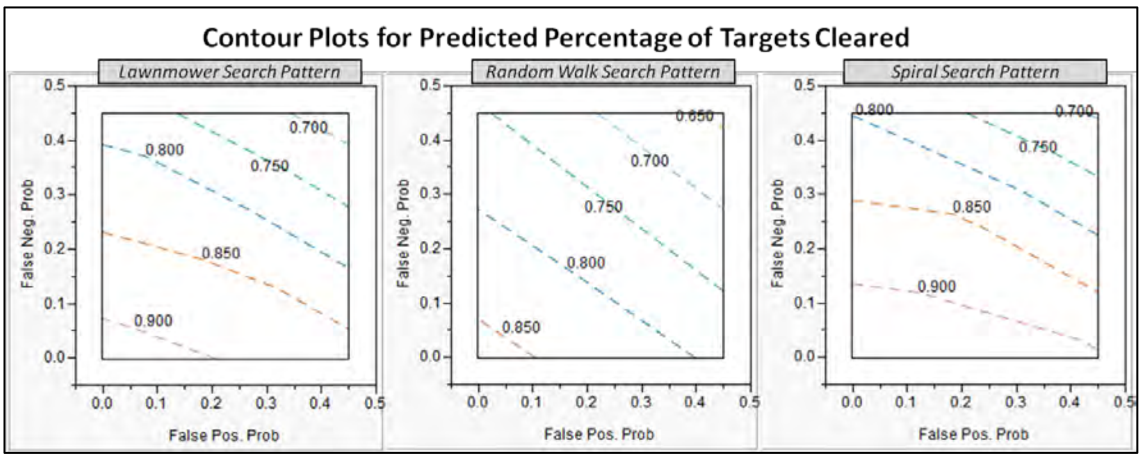


Figure 11. Surveyor UAV range of sensor characteristics, all search pattern

Let us look at another example of studying the data to determine which scenarios achieve an 85% target clearance rate. From the data collected in the DT phase, we built a prediction profile (Figure 12) of the relevant factors, providing a graphical illustration of how they interact. The vertical axis presents the logit response when we select different factor settings. Recall from Equation (7) that a logit response of ~1.7346 results in an 85% target capture rate, and any larger logit value results in a higher percentage of targets captured. In Figure 12, we can see that the Spiral search pattern outperforms Random Walk, and increasing false negative and false positive probability decreases the target clearance percentage. This confirms the results shown in Figure 11.

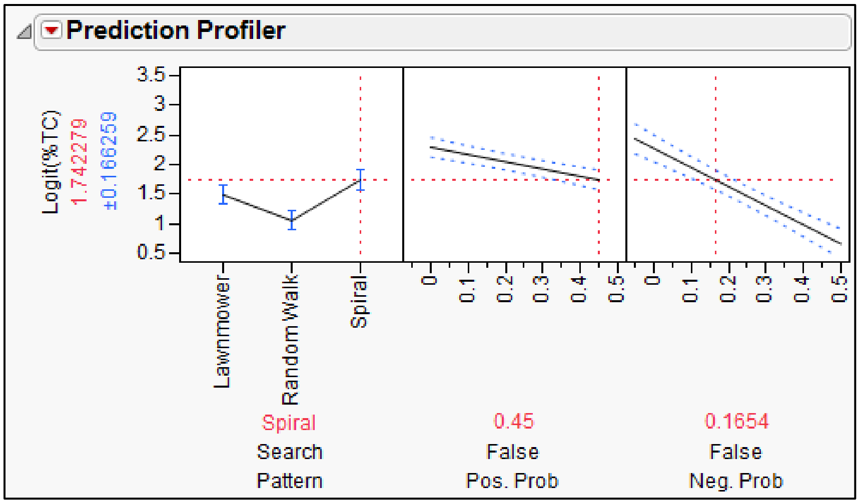


Figure 12. Prediction profile of percentage of targets cleared from the DT phase

The results of the DT DOE show which factors are significant, how they affect the response, and what values of γ and ρ satisfy the 85% requirement. Various factors come into play when the design program office is selecting which design parameters to present for OT. Some examples include production and development costs, physical engineering limitations, and operational limitations and constraints. The DOE Analysis phase allows us to conduct a sensitivity analysis to investigate the range of options that best suit the criteria required by the SUT and by the design team. We see this graphically in Figure 13.

Having selected Spiral search as the most effective pattern, we build a contour profile (Figure 13) that effectively presents a sensitivity analysis of false positive probability (γ) vs. false negative probability (ρ). We set the contour slider to 1.7346 to attain the desired target capture rate of 85%. Factor level combinations of γ and ρ intersecting above the contour line in the shaded area violate the desired response criteria. SBT&E looks for a specific combination of key performance parameters at this point. CBT&E, however, inherently provides a wide range of suitable combinations.

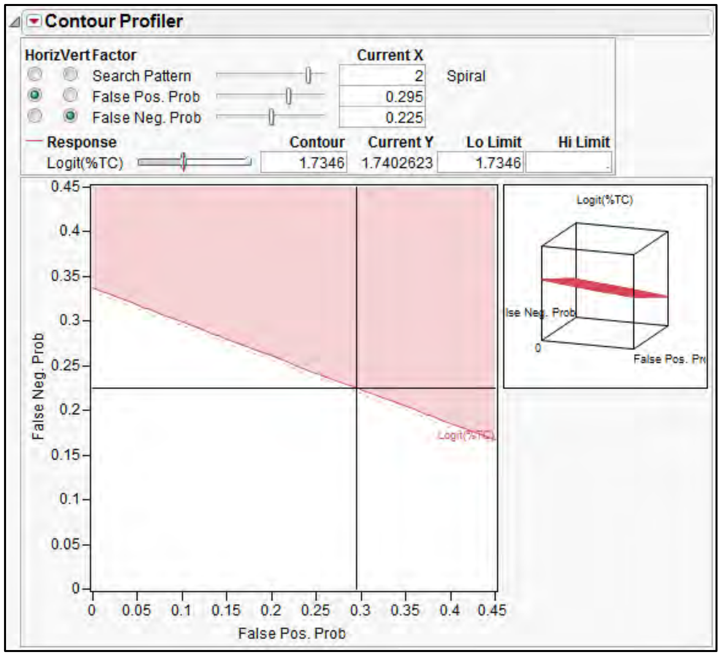


Figure 13. Contour presentation of γ vs. ρ as a function of search pattern and desired response

2. Operational Testing (OT) Phase

The second phase of our analysis deals with the operational concerns of the end user, where all factors in the model are variable, but with Surveyor UAV factors set to specific factor levels as learned in the previous testing phases. Operational employment considerations become relevant as we allow for expansion of the AOI, the number of hostile targets, and variation of the levels of system integration between Surveyor UAV, Tracker UAV, and the QRF.

a. Planning and Design Considerations in OT

As part of the continuous process of the DOE conceptual cycle presented in Figure 7, test authorities must again spend considerable focus of attention on the planning phase of DOE. The fundamental difference in this process now, however, is the incorporation of information collected from the DT phase. We plan our designs using the analysis from DT to guide the implementation of system integration in OT.

For example, in our case, the objectives (MOE and MOP) of the test remain the same, to capture at least 85% of the hostile targets that ingress the AOI, but the factor space expands to include all of the factors identified in Table 2. Stakeholders in the test have to select particular settings for γ, ρ and Search Pattern to continue the test. They also need to decide which aspects of the SUT are available for compromise in order to support other design considerations and still meet capability requirements. Operationally, in the absence of other factors, lower false negative probabilities are preferred, but there may be a bottom limit (lower bound) that is prohibitively more costly to attain in terms of engineering, time or financial expenditure. Conversely, a high false positive probability and low false negative probability may be easy to engineer in Surveyor UAV, but the increased number of targets to investigate may negatively influence the performance of the QRF and SoS as a whole.

We used the results from DT to fix factor settings for false positive rate (γ), false negative rate (ρ), and Search Pattern for T&E in the OT phase. In the next section, we examine four different perspectives that still meet the design specifications:

1. Lowest (best) False Negative rate
2. Highest Suitable (worst) False Negative rate
3. Mid-Range False Negative rate
4. Better-than-Specifications design (~90% target capture rate)

The presented perspectives could each represent valid design, systems engineering or budgeting concerns of the design program office that affect production of the SUT.

In our design phase, we have selected the four scenarios presented in Table 5 for our factor settings and levels. We hold the Surveyor UAV characteristics constant at the levels determined in the DT phase (an SBT&E approach), and set Tracker UAV factors levels as indicated. We

developed the hold-constant factors (γ, ρ, and Search Pattern) of our four designs by using the contour profile presented in Figure 13. Design A is the best false negative case that lies on the 85% target capture contour line. Design B is the point on the contour line where false negative probability is at a midpoint, and Design C is where we eliminate any false positive probability rate. Finally, we selected Design D to represent a point where the Surveyor UAV SoS should exceed capability requirements and capture ~90% of inbound hostile targets on the DT test range.

	Design A	Design B	Design C	Design D
False Neg. Prob (ρ)	0.45	0.225	0.335	0.14
False Pos. Prob (γ)	0.165	0.295	0	0.14
Search Pattern	Spiral	Spiral	Spiral	Spiral
Tracker Launch	5	5	5	5
Tracker Speed	3	3	3	3

Table 5. Factor Levels for those factors held constant during the OT phase

The variable factors consist of two categorical factors (one with two levels and one with three) and four continuous factors. A full-factorial design, like that presented in the DT phase, would require $2 \times 3 \times 2^4 = 96$ design points. In order to detect quadratic effects, we would need to augment the design with center points, requiring an additional $2 \times 3 \times 4 = 24$ design points. In the real world, a design requiring 120 individual design points is likely cost and time prohibitive. We selected a D-optimal design for main effects, two-factor interactions and quadratic terms encompassing 48 design points to examine the performance of the Surveyor/Tracker/QRF SoS. Optimal designs are those that allow analysts to select an appropriate design based on a hypothesized regression model. They offer advantages in DOE by reducing the cost of T&E by reducing the number of experimental trials, and being able to accommodate multiple types of factors. D-optimality minimizes the variance of the regressor

coefficients (Montgomery, 2009). This is a good choice for designs with larger numbers of factors, such as those encountered in OT.

We present our design in Table 6. The D-optimality includes mid-range values for each factor level, rather than just the endpoints of the factor space. This is an important feature of our design because it allows us to estimate any quadratic effects present in the model. We expect quadratic effects to be significant in the Search Area factor. For instance, as we double the edge lengths of our AOI from 10 kilometers to 20 kilometers, the size of the area actually quadruples (from 100 sq km to 400 sq km). This affects not only the size of area that Surveyor and/or Tracker UAV must cover, but also influences the *target density,* defined as the number of objects per square kilometer (i.e., 30 objects in a 10x10 km AOI results in a target density of 0.3).

Design Point	Team Type	Object Motion	Interdictor Transit Time	Clear Time	Number of Objects	Search Area
5	Surveyor	FastRW	15	1	90	2500
6			15	21	30	2500
8			1	21	90	100
12			1	1	30	2500
21			1	21	30	2500
29			1	1	60	100
30			15	1	30	100
38			8	11	90	1296
7		SlowRW	1	21	90	2500
10			15	1	30	2500
20			8	1	30	1296
23			15	1	90	100
27			15	21	90	2500
43			1	21	30	100
46			1	1	90	100
47			1	11	60	2500
48			15	21	60	100

Design Point	Team Type	Object Motion	Interdictor Transit Time	Clear Time	Number of Objects	Search Area
9	Surveyor w/ Tracker UAV	FastRW	1	11	30	100
15			15	21	90	100
17			1	21	90	2500
33			15	1	30	2500
35			15	21	30	100
41			1	21	60	1296
42			15	1	90	2500
44			1	1	90	100
1		SlowRW	1	1	90	2500
2			15	1	90	1296
4			15	1	30	100
11			1	21	90	100
13			1	1	30	2500
18			15	21	30	2500
22			1	21	30	100
40			15	21	90	2500

Design Point	Team Type	Object Motion	Interdictor Transit Time	Clear Time	Number of Objects	Search Area
14	Surveyor w/ Tracking Mode	FastRW	1	11	90	2500
19			1	21	30	100
26			8	1	30	2500
28			15	1	90	100
31			1	1	90	1296
34			15	11	30	1296
39			15	21	90	2500
45			8	21	60	100
3		SlowRW	15	1	60	2500
16			8	1	90	2500
24			8	11	90	100
25			1	1	30	100
32			1	21	30	2500
36			15	21	30	100
37			1	21	90	1296

Table 6. OT Phase Experimental Design, grouped by categorical factors

b. Execution and Analysis of OT Results

Once the experiments were completed, we moved directly into the analysis portion of OT. We began with a straightforward descriptive analysis of the response for each of the four design scenarios. Table 7 presents the percentage of responses across the 48 design points of each design that achieve the specified target capture rate (greater than 50%, 60%, etc.). For example, one of 48 design points (2.08%) in design A resulted in better than an 85% target capture rate. Note that these are cumulative in nature, and not binned within percentage bands. For these designs, all Surveyor/Tracker/QRF teams fail to achieve the desired objective.

Percentage of OT Design Points Achieving Specified Target Capture Rate					
Design	≥ 50%	≥ 60%	≥ 70%	≥ 80%	≥ 85%
A	10.42%	6.25%	4.17%	2.08%	2.08%
B	10.42%	8.33%	4.17%	2.08%	2.08%
C	10.42%	10.42%	8.33%	6.25%	2.08%
D	10.42%	10.42%	6.25%	4.17%	0.00%

Table 7. Percentage of OT Design Points by Target Capture Rate

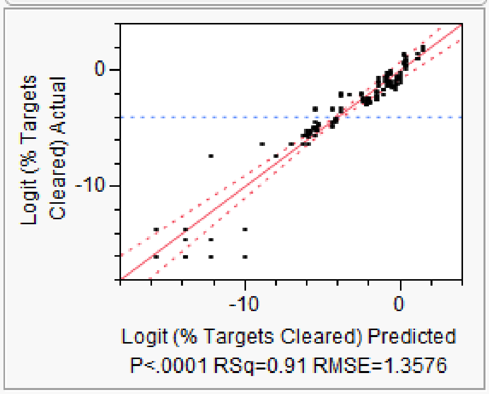
It is apparent from the basic descriptive statistical analysis that if this were an actual OT&E evolution with the stated evaluation criteria (a minimum 85% target clearance rate), our SUT is not operationally effective or operationally suitable. This is consistent with the 2008 charter and subsequent findings of the Defense Science Board. Specifically, "approximately 50 percent of programs entering IOT&E in recent years have not been evaluated as Operationally Effective and Operationally Suitable" (Defense Science Board Task Force, 2008). Our challenge is to analytically examine the program data and determine the root causes of failure in this instance.

To conduct further study, we aggregate the data collected across all four of the design scenarios and analyze it as a single test. We present initial

fitted model in Figure 14. While the Summary of Fit statistics appear satisfactory, observation of the actual by predicted plot illustrates a significant departure of data points at the lower left corner. This indicates problems with certain assumptions of model validity required in statistical analysis. In particular, the error terms (called residuals) should exhibit constant variance across all design points. This leads us to reject the initial model.

Response Logit (% Targets Cleared)

Actual by Predicted Plot



Summary of Fit

RSquare	0.909674
RSquare Adj	0.897308
Root Mean Square Error	1.357588
Mean of Response	-4.02631
Observations (or Sum Wgts)	192

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	23	3118.3071	135.579	73.5622
Error	168	309.6318	1.843	**Prob > F**
C. Total	191	3427.9389		<.0001*

Figure 14. Combined Model of all OT Design Scenarios Summary of Fit and ANOVA

A deeper analysis of the data, however, leads us to a particularly insightful observation. Observe the fitted percentage of targets cleared against Search Area grouped by Team Type as presented in Figure 15. The curves depicted represent the percentage of targets cleared in our fitted model as a function of Search Area, but grouped by Team type. There is a noticeable difference in the performance of Surveyor UAV only (no tracking capability)

against Search Area. From this, we chose to re-fit a model on the combined data, but excluding observations involving Surveyor only.

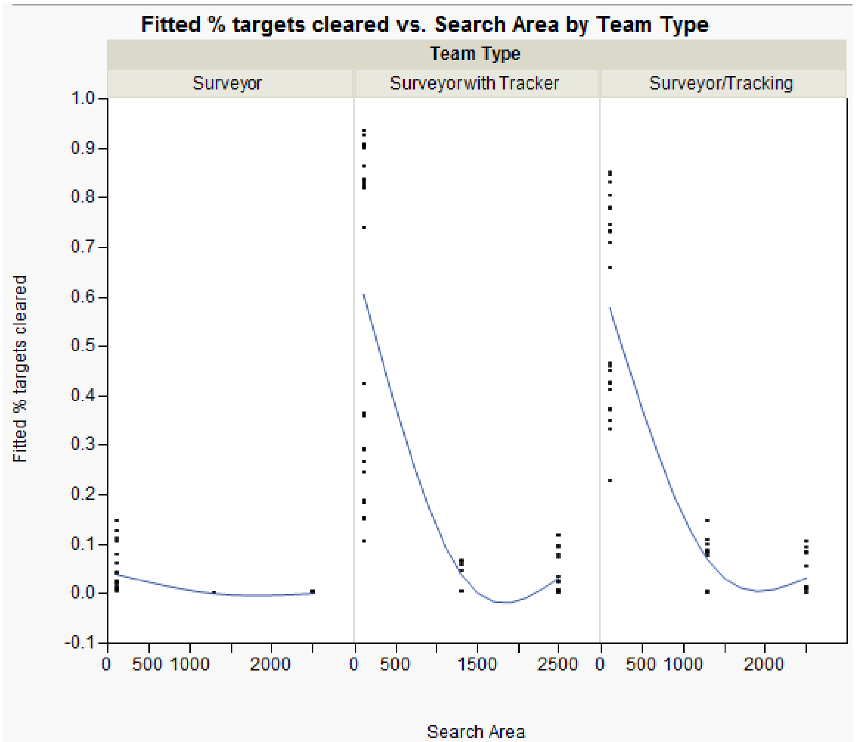
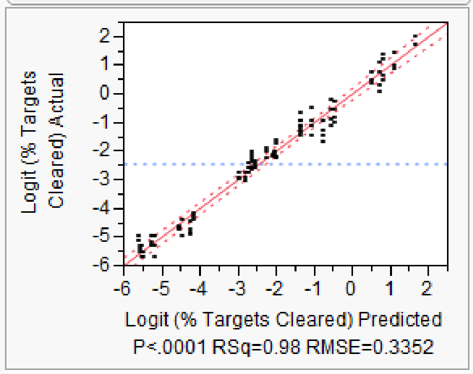


Figure 15. Combined OT model, MOE vs Search Area grouped by Team Type

This resulted in a much cleaner model that satisfied the necessary assumptions. Figure 16 shows much improved performance metrics in R^2 and R^2-adj, as well as improved accuracy of the actual by predicted plot. Using this fitted model, it is much simpler to determine the most significant factors affecting our MOE. Observation of the parameter estimates confirm our intuition that the greatest effects on SUT performance come from Search Area, Interdictor Transit Time, Number of Objects and their associated two-factor interactions. Additionally, Search Area and Clear Time quadratic terms were also significant to this model.

Actual by Predicted Plot



Summary of Fit

RSquare	0.979819
RSquare Adj	0.976359
Root Mean Square Error	0.335187
Mean of Response	-2.44696
Observations (or Sum Wgts)	124

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	18	572.75248	31.8196	283.2169
Error	105	11.79681	0.1124	**Prob > F**
C. Total	123	584.54929		<.0001*

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
				9.4962
Lack Of Fit	12	6.495621	0.541302	**Prob > F**
Pure Error	93	5.301190	0.057002	<.0001*
Total Error	105	11.796811		**Max RSq**
				0.9909

Figure 16. Combined Model Summary of Fit, excluding Surveyor only

Furthermore, one additional technique that proves useful to the analyst is the utilization of a partition tree to map the effects of various factor level settings in our model. Presented in Figure 17 and available in most statistical software packages, it uses a method known as recursive partitioning to split the source data into subsets grouped by attribute values and create a predictive value for each subset (based on groupings of factors that best predict a response value). The software continually partitions each subset of data until it can extract no more value from the data (SAS Institute Inc., 2011). From this partition tree, we observe that the greatest performance from our SUT comes under conditions where we limit Search Area to less than 1296 square

kilometers, we exclude Surveyor only from the Team Type categorical factor, and we hold Interdictor Transit Time to less than eight time steps.

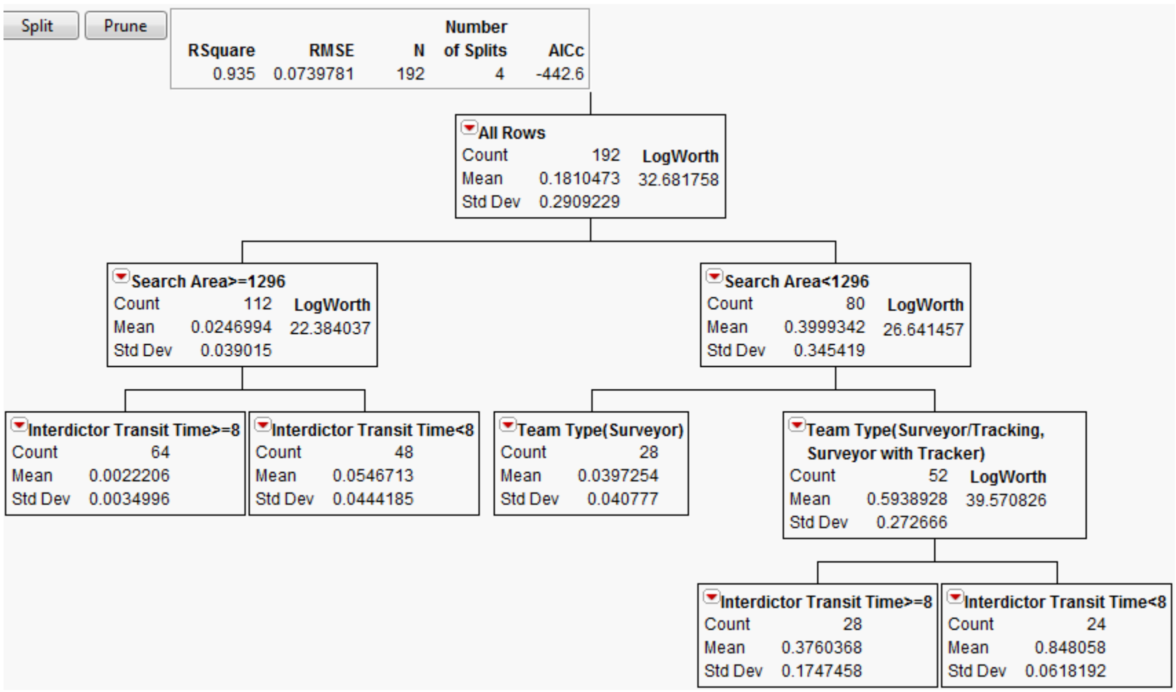


Figure 17. Partition Tree on Combined OT data showing conclusions regarding factor level value

The benefit of this analysis is that it provides valuable insight as we reanalyze performance of the SUT from an IT perspective. Prior collaboration between program offices in earlier in the T&E process would have enabled recognition of the limitations of this SUT. By combining prior collaboration with rigorous statistical studies, it is likely that we would have found the majority of the discrepancies in this particular OT program much earlier and at much less cost. As stated by the Defense Science Board, “operational influence and perspective earlier in the developmental process is a proven catalyst for early identification and correction of problems” (Defense Science Board Task Force, 2008).

3. Integrated Testing (IT) Phase

The IT phase deals with the expansion and modification of T&E to include integrated testing, where considerations of the Surveyor UAV designer and the

end user (tactical operator) share common interests across organizational lines. The IT phase collaborates and shares data in support of independent analysis, evaluation and reporting by all stakeholders. In this way, we are able to validate the performance of our methodology and its applicability to CBT&E.

In this section, we show how lessons learned from a collaborative process in DT and OT result in better overall system performance. We recognize that the decision to fix γ, ρ, and Search Pattern at fixed levels from the DT phase for operational testing may have been ill-advised considering the lack of operational input within the DT phase. We can generally consider this practice to be one of the root limitations of the SBT&E design philosophy. Proper IT is requires contractor and operator collaboration. Furthermore, the examples we present in this paper generally apply when one expands consideration of this methodology from our *notional* example T&E scenario to actual MDAP and T&E programs. Additionally, we take the opportunity to highlight some of the advantages gained when we incorporate M&S techniques within the CBT&E process.

For discussion purposes only, we treat IT as an independent and stand-alone testing phase (as discussed in Chapter II). In actuality, IT should exist as a continually updating and repeatable process spanning both DT and OT. Although we have already presented the detrimental results of OT in this paper, we now conduct and analyze IT experimentation as if it were an intermediate step bridging the gap. We do this to confirm OT results and highlight where we can gain efficiencies much earlier in the T&E process.

a. Planning and Design Considerations in IT

We again turn to the conceptual cycle of DOE presented in Figure 7 to frame our discussion of the IT phase results. Comprehensive planning is the first step (and perhaps most critical) in our methodology, for this is an excellent time to capitalize on opportunity cost savings across time, risk, and budgetary concerns. Stakeholders in IT, including both system-level engineers and mission-level operators need to consider the overall scope of the problem in

order to identify areas where resource sharing is most effective. As Dr. Streilein of ATEC wrote:

> The T&E strategy must do more than check a system's capabilities against the standard type of requirements; now the mission capabilities must also be outlined and a crosswalk developed to ensure that the test events and data will address both system and mission capabilities. (Streilein, 2009)

Planners should also recognize the positive impact that M&S can have on the IT process. The complexity of the operational environment makes it infeasible to test every possible mission scenario, or offer a sufficient number of replications or observations to attain the appropriate statistical significance. However, M&S tools, such as SASIO or other verified, validated and accredited simulation models, do provide methods in lieu of live testing for program managers and contractors to enhance system design. The DSB findings state, "most developmental and operational tests should be preceded by M&S to predict test outcomes, with corrections to models and data made as required following a block of testing" (Defense Science Board Task Force, 2008).

The following design phase consideration and example highlights the utility of M&S. One objective of DOE in T&E is factor screening, a process by which we vary the input factors to determine which are most influential on the response variables. This screening includes the main effects as well as any interactions between factors. Systematically changing factor levels and observing the effect of the response is what enables us to model mathematically the process under test.

Two general design categories useful for screening are full factorial and fractional factorial designs. A full-factorial design is a basic form of exploring the factor space, in which the experimenter examines every relevant factor level against every other combination of relevant factor levels. Although we obtain very complete and detailed data this way, the large number of experiments required makes this method inefficient and undesirable. For example, in an experiment involving two three-level factors and nine two-level factors, the

number of experiments required is $3^2 x 2^9 = 4{,}608$ design points with zero replications. A large number of design points requires a large budget in time and resources; our example is likely cost and resource prohibitive in the T&E environment, but perfectly feasible in an M&S environment.

A more efficient method of factor screening is by using fractional factorial design to examine the main effects and second-order interactions only. With this method, we are looking to identify the factors that have large effects on SUT performance. Additionally, we reasonably assume that higher order effects (e.g., three factor or higher interactions) are negligible. This type of design leverages the "Effect Sparsity Principle," which states, "The number of relatively important effects in a factorial experiment is small" (Wu & Hamada, 2000). Subsequently, we can use significantly fewer experiments (at a much lower cost) to gain important information on main effects and low-order interactions. We then use subsequent experiments (such as augmentation for quadratic effects, as budget constraints allow) to investigate the most important factors in more detail.

Optimal designs, as presented previously in the OT phase, are a special case of designs that also offer significant advantages in T&E. For this IT phase example, we used a D-optimal design for main effects, two-factor interactions, and quadratic effects in Search Area encompassing 96 design points across the 11 input factors. While 96 design points might seem expensive, it represents more than an order of magnitude improvement over 4,608 design points.

b. Execution and Analysis of IT Results

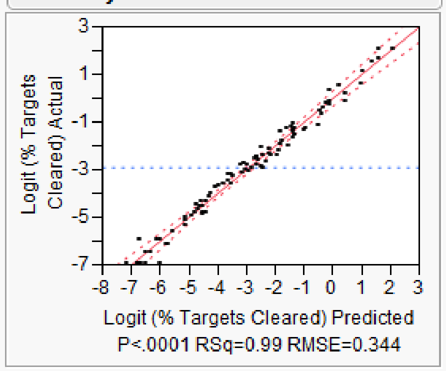
In IT, we need to undertake test plan execution with the special consideration that operators conduct events in order to ensure that they preserve the independence of data collected for use in OT analysis. This is in accordance with the requirements of U.S. Title 10 code outlining the legalities of OT&E

(10USC2399, 2002). Other than that concern, we treat the collection and analysis of IT data in the same manner as previously demonstrated.

For this example IT program, we consider ourselves much earlier in the overall TEMP. Following execution of our D-optimal design, we develop a model that accurately predicts the actual behavior of the observed test articles. For a large design, the number of combinations of regression coefficients is generally too large to allow for explicit examination of all possible subset combinations. Thus, we utilize a technique called stepwise regression, which uses statistical software automation to search the large factor space for the best predictive combination of regression coefficients. From a relatively small number of design points we obtain a model that adequately predicts the response variable. We present summary statistics of our fitted model in Figure 18.

Response Logit (% Targets Cleared)

Actual by Predicted Plot



Summary of Fit

RSquare	0.98506
RSquare Adj	0.978165
Root Mean Square Error	0.343999
Mean of Response	-2.90901
Observations (or Sum Wgts)	96

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	30	507.15729	16.9052	142.8588
Error	65	7.69180	0.1183	**Prob > F**
* C. Total	95	514.84908		<.0001*

Figure 18. Summary data for IT phase predictive model.

We observe good parametric data in R^2-adj and Root Mean Square Error. We use this model for factor screening, which enables to identify which

factors and factor combinations exhibit significant effect on the percentage of targets cleared. In this case, the model indicated significant negative factor effects caused by the employment of Surveyor only (no tracking capability), and increases in Search Area, Interdictor Transit Time, False Positive and False Negative probabilities. Slightly positive factor effects came from the increasing the Tracker UAV launch distance and when Object Motion slowed. Multiple two-factor interactions also proved statistically significant. From these observations, we obtain both systems engineering and operational insights that reset designer expectations.

The poor performance of the Surveyor/Tracker FoS with respect to the MOE presents cause for concern. The additional complexity of this controlled operational environment proves detrimental to our SUT performance. The capability we were trying to meet with this SUT was to capture at least 85% of the hostile targets that ingress the AOI. However, interim analysis accomplished within the IT phase makes it apparent that with the existing conditions this goal is likely too ambitious. On the positive side, though, catching this error earlier within the T&E process allows changes and/or re-design to be accomplished in a more timely and cost effective manner.

To illustrate this more clearly, in Figure 19 we present selected contour profiles that show the limited performance. Each profile shows the various combinations of γ and ρ that attain the average percentage of targets cleared. The contour profile in the upper left corner clearly demonstrates the best overall performance, but still only attains an average 70% of targets cleared. As the complexity of the operational environment increases (i.e., longer transit time and clear time, greater search area), the worse the overall system performance becomes. In contrast, the bottom right contour profile shows an accomplishment of only 10% targets cleared.

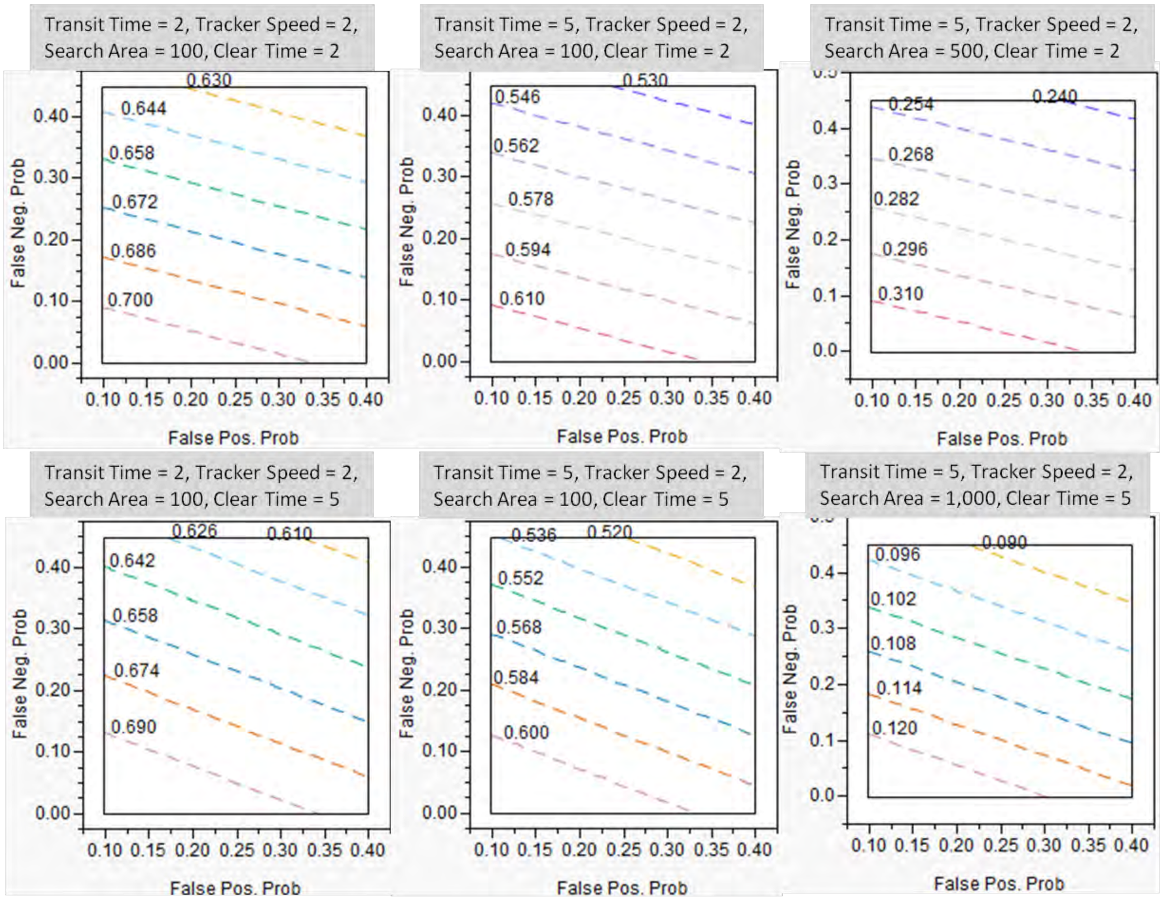


Figure 19. Selected Percentage of Targets Cleared as function of the Sensor Performance Parameters, demonstrating declining performance

SUT performance is indirectly proportional to search area size; in the same fashion, it is inversely proportional to QRF clear time and transit time. From an OT perspective, it is important to note that search area and transit times are considerations of operational employment tactics, and clear time is a function of QRF training. Evaluators should address operational as well as engineering concerns in an integrated fashion. These relationships lead us to look for factor constraints (like search area size limitations), system engineering level factor improvements, or operational doctrine employment strategies to meet capability requirements. In certain cases, re-evaluation of the programmed SUT capability requirements may be the only solution.

It is also important to recognize in the IT phase the possibility of interactions between input factors adversely affecting the performance of the SUT. In the DT phase, we conducted our experiments in a controlled environment, whether in the laboratory or under test range conditions selectable by the design authority. From these sterile conditions, we selected target levels for false positive (γ) and false negative (ρ) probability values and fixed them as a system engineering consideration. However, Figure 20 illustrates two situations in which interactions between Surveyor UAV sensor characteristics and the QRF (interdictor) performance characteristics exist. In the first plot, we see that when γ is fixed at 0.0, there is no change in the observed MOE. However, when γ is fixed at 0.45, when Interdictor Clear Time is increased SUT performance decreases. Likewise in the second plot, regardless of the setting for ρ, SUT performance decreases with an increase in Interdictor Transit Time. However, the effect is more dramatic with ρ = 0.0 than it is with ρ = 0.45. In both cases, γ and ρ at the 0.0 factor level dominates the 0.45 factor level with changes in clear time or transit time. We miss the effect of these interactions in an SBT&E environment.

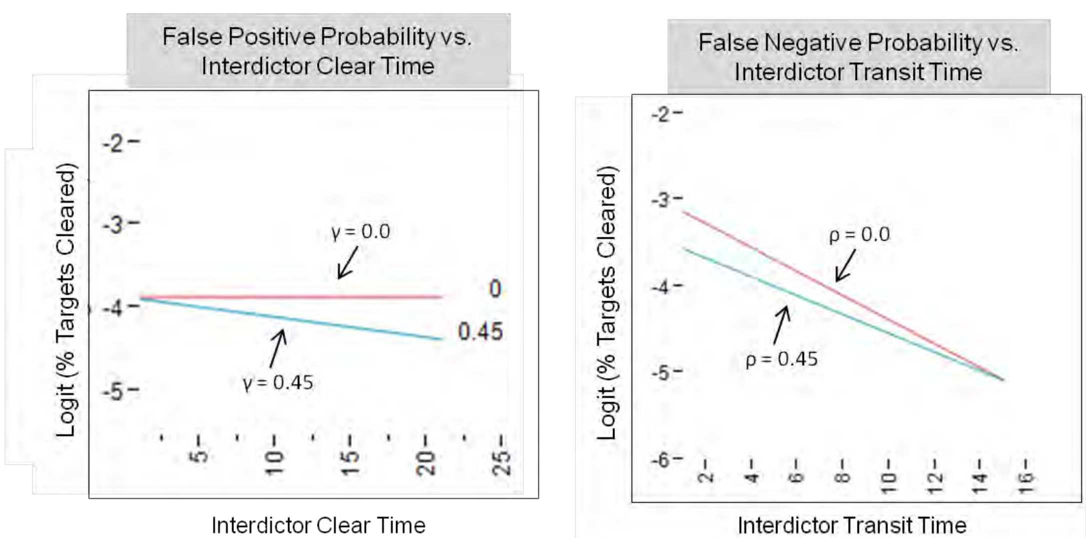


Figure 20. Surveyor UAV sensor characteristics vs QRF performance characteristics interaction plots from OT phase predictive model

Using a combination of recursive partitioning techniques and examination of contour profile tools, we observed the best system performance when Search Area was limited to less than 1,296 square kilometers, Interdictor Transit Time was less than eight time steps, and the Teaming Type contained some form of tracking capability. These are all factors that are capable of modification or constraint by the operator. From an engineering perspective, Surveyor false detection probabilities were significant; however, preferred values never fell below 0.10. Finally, uncontrollable factors like the Number of Objects and Object Motion characteristics held significance, and due to the partially controlled nature of IT, we felt comfortable limiting these factor levels for the purposes of factor space exploration. This led us to plan additional experimentation (which we call Design E), as specified in Table 8.

Test Factors	Factor Levels
False Positive Probability (γ)	[0.1, 0.45]
False Negative Probability (ρ)	[0.1, 0.45]
Search Pattern	[Spiral, Lawnmower]
Tracker Launch	[1, 5]
Interdictor Transit Time	[8, 1]
Tracker Speed	[1, 3]
Search Area	[100, 1296]
Clear Time	[1, 11]
Number of Objects	[30, 60]

Held Constant Factors	Setting
Object Motion	SlowRW
Team Type	Surveyor with Tracker

Table 8. Redesign Parameters for IT phase sequential test plan (Design E)

Thus, analysis led to planning, and planning led to re-design, completing an entire circuit of the conceptual cycle for experimental design. Based on our re-evaluation of the factor screening observations, we limited factor

levels as specified and established a D-optimal design in main effects, two-factor interactions, and quadratic effects in search area with only an additional 64 design points.

Table 9 presents a basic descriptive analysis of the response for the re-design scenario, just a presented in Table 5. While significantly below the required design criteria, there is definite performance improvement under the new test conditions.

Percentage of IT Re-Design Points Achieving Specified Target Capture Rate					
Design	≥ 50%	≥ 60%	≥ 70%	≥ 80%	≥ 85%
E	26.56%	18.75%	12.50%	3.13%	1.56%

Table 9. Percentage of IT Re-Design Points by Target Capture Rate

It is important to reiterate at this point the power of incorporating M&S as an integral part of the IT process. Effective M&S tools that accurately model system performance ease potential burdens encounter with multiple design point requirements. Sequential design of this nature could be useful for discovering the proper factor settings or superior performance regimes.

C. ANALYSIS SUMMARY

Throughout this chapter, we have presented a flexible methodology for incorporating DOE and M&S into the T&E process. The methodology is flexible in the sense that a test authority can:

- Choose from a number of different experimental designs depending upon the objectives of the particular test regime;
- Perform many different analyses of the same dataset using a myriad of powerful statistical tools;
- Discover a great deal of information about the SUT, whether intended or unintended, that might prove beneficial to the T&E process;

- Realize a wide variety of time, cost and risk savings early and upfront in the T&E Master Plan when changes have the greatest impact for the least cost.

These tools represent a small subset of analytical techniques that greatly enhance a test plan developer's "tool kit." These advanced tools are useful in capturing and analyzing data over the life of a system, and not just during the initial phases of development and design.

We have presented one method of conducting DOE across a range of input factors. We use DOE to study how changing the levels of independent input factors affect the overall variability in a model. It is important that test plan designers avoid a single-minded focus on particular specifications rather than a range of capabilities. This is not to say that we disregard the achievement of key performance parameters and critical requirements. It is simply a means of focusing on the big picture in lieu of the small.

In today's operationally diverse military environment, T&E activities can no longer afford to operate under the SBT&E construct used in the past. Warfare has evolved, requiring our military operators and systems to evolve with it. The Acquisition process cannot afford to find itself struggling to field operationally relevant systems. CBT&E adopts flexibility and robust design philosophy intended to capture the wide range of capabilities necessary for the modern warfare environment.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSIONS AND RECOMMENDATIONS

At the onset of this research, we set out to accomplish two primary objectives:

- Illustrate the positive effect of incorporating DOE and M&S techniques throughout the entire T&E process
- Quantitatively demonstrate the benefits of CBT&E over SBT&E.

In this chapter, we summarize our results, and explore possible time, cost and risk savings through utilization of systematic analytical methodologies in conjunction with proven statistical techniques. We provide recommendations for future work in this area to enhance and streamline the T&E process.

A. EXPLORING THE DOE METHODOLOGY

In a November 2010 briefing to NPS students and faculty, Dr. Catherine Warner, Science Advisor to the Director, Operational Test & Evaluation Command, stated, “No ‘one size fits all’ approach exists when applying DOE in defense acquisition test and evaluation.” Our research certainly exemplifies this statement, as test authorities will need to individually and specifically tailor their T&E master plans to the systems under test. However, we have demonstrated through illustrative example that one can modify a wide variety of standard techniques and commonly used designs to field relevant systems at reduced cost.

We have presented the design objective known as factor screening, which uses designs like factorial, fractional factorial, and D-optimal designs to achieve specific results. Additional design objectives that we have not discussed, such as response surface methodology and robust design, utilize different DOE techniques to examine alternative facets of the SUT. Valid design approaches include other optimal design variants, Taguchi methods, Plackett-Burman designs, and space-filling designs. M&S opens design availability even more.

Furthermore, augmentation and sequential design techniques using basic DOE provide an easy method of meeting the specific requirements of any given situation.

Continuing research in DOE presents new opportunities. Development of a methodical master plan and complete testing strategy to accomplish capability objectives is critical. Application of the conceptual cycle of experimental design is a systematic philosophy useful in concentrating the proper focus of effort in all phases, DT, OT, and IT, of CBT&E.

B. EMPHASIZING MODELING AND SIMULATION IN ALL T&E PHASES

By using a simulation model as a proxy for an actual test evolution, we have also demonstrated the advantages of incorporating the power of M&S to inform decision-makers and enhance system performance. The original purpose of the SASIO model was to act as a modeling framework to aid ISR operators gain insight on tactical employment techniques. We borrowed its capabilities to demonstrate the utility of simulation as a design tool in the T&E environment. Fully validated, verified, and accredited models currently in use, such as STORM and BRAWLER, provide a more robust ability to examine the full range of mission scenarios across an extremely large factor space. This enables system engineers and operational planners to determine capability areas truly important to the war fighter, and thus constrain costly T&E efforts to that which is most important.

Furthermore, computer-aided design enhances the ability of designers to fully explore a myriad of design and employment options that were not possible in times past. The accessibility of extremely capable computing power, either on standalone super-computers or on clustered networks applying computational power in parallel, provides a great opportunity to investigate options previously denied because of excessive risk or cost. Computational power simulating the real world is relatively inexpensive in comparison to live events.

C. CAPABILITIES VS. SPECIFICATIONS BASED T&E COMPARISON

We have emphasized DOE and M&S as tools critical to the development of the CBT&E processes. We have shown analytically the advantages of flexible, robust methodologies in the development of T&E master plans. Identification of the most important variables of the process under test is a critical first step that rigorous and structured testing can help accomplish. Additionally, the systematic application of the Plan-Design-Execute-Test cycle of DOE often results in identifying factors previously overlooked under the SBT&E concept. Rather than learning of potential setbacks late in the T&E process, such as in OT evaluations, we incorporate a flexible yet structured process during all phases of design and execution.

The IT phase that we have demonstrated in this thesis serves as an effective tool in the completion of the T&E process. As we strive to shorten acquisition timelines while meeting performance and cost requirements, IT assists in achieving shared efficiencies between government and contractor personnel. In fact, DoD Instruction 5000.2, as well as by direction of the Undersecretary of Defense (AT&L) and DOT&E have mandated the use of integrated testing in T&E (Defense Science Board Task Force, 2008). This effectively allows us the opportunity to identify and modify factors influential to the SUT much earlier in the design process.

D. ONGOING AND FUTURE WORK

Ongoing efforts by the NAVAIR CBTE Working Group continue to explore methods of ensuring delivery of the right Integrated Warfighting Capabilities (IWC) to Navy operators. This effort serves to modify analysis from a one-time, up-front process to a primarily continuous process consistent with the experimental design cycle. Concurrent work by the U.S. Air Force in Capabilities Based Evaluation and by the U.S. Army with Mission Based Test Design is also underway.

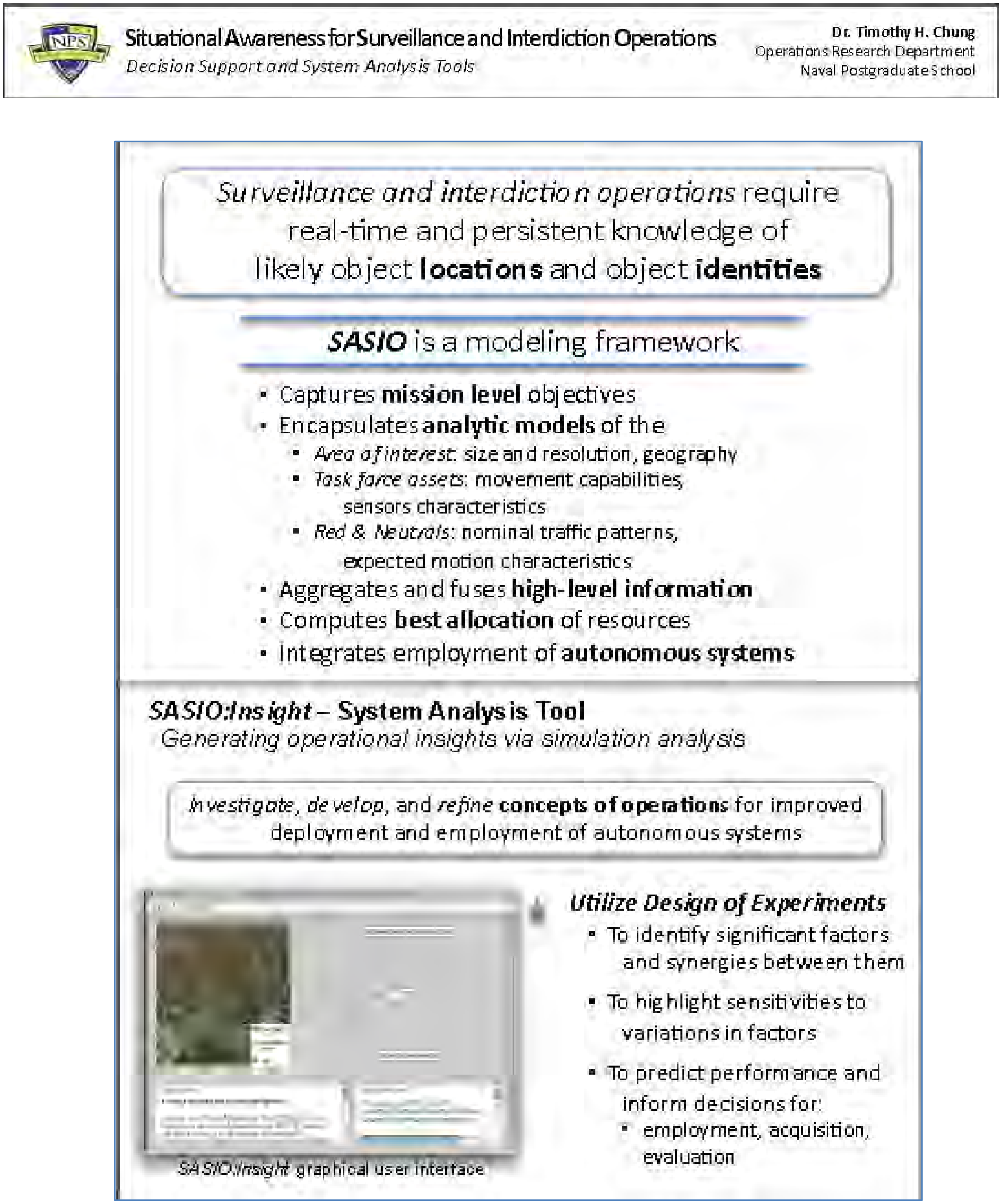
Evaluators in all services have been exploring the utilization of Live, Virtual, and Constructive (LVC) testing, sometimes referred to as distributed network testing, to evaluate the performance of SoS constructs where assets are distributed at various locations worldwide, but interconnected by secure Virtual Private Networks (VPNs). Development of a simulation/optimization support tool to determine the optimal allocation of flight/ground testing vs. distributed network testing to minimize time, risk and budgetary cost would be useful. Along the same lines, a cost-based analysis regarding the level of savings available in the same functional areas through elimination of certain live test events in favor of distributed network-based sharing capabilities would provide quantifiable metrics for CBT&E implementation.

Future research opportunities building on this work could support CBT&E in the following ways:

- Exploration of sequential design and design augmentation techniques in support of specific T&E goals; and
- Exploration of the combination of live experimentation with simulation experimentation, and its impact on the T&E process.

Additionally, each Service Operational Test Authority has different processes, procedures and approaches to the capabilities-based planning effort. Further work promoting the standardization of T&E and Acquisition processes from the perspective of the Joint force would enhance the future integration of military mission systems. Many more avenues in this field of work exist for the interested researcher. Improvement of the T&E process is a continually evolving area of study.

APPENDIX A – SASIO SIMULATION TOOL



SASIO:Command – Decision Support Tool

Providing decision recommendations for improved effectiveness

Better information gathering

- Integrate multiple data, sensor, and intelligence sources
- Employ multiple distributed, networked, autonomous assets
- Quantify probabilistic uncertainty in model and environment

... leads to ***better decision making***

- Enhanced routing of search assets
- Enhanced task allocation of heterogeneous assets
- Enhanced courses of action in response to dynamic situations

SASIO:Command
graphical user interface

Bridging Autonomous Systems and Operations Research

Integrating theory and experimentation for future concepts

Current Experimentation Efforts

- Active quarterly participation in the **NPS-USSOCOM Field Experimentation Cooperative** (TNT) at Camp Roberts, California
- Live-fly experiments for **Joint Expeditionary Force Experiment** (JEFX) in support of Second Fleet's Maritime Operations Center missions

New technologies enable new operational concepts
New operational concepts drive new technologies

Laboratory for Autonomous Systems and Operations Research

Email: *thchung@nps.edu* ***Web***: *http://faculty.nps.edu/thchung*

APPENDIX B – SERVICE OTA MEMORANDUM OF AGREEMENT

MEMORANDUM OF AGREEMENT

SUBJECT: Using Design of Experiments for Operational Test and Evaluation

Regarding the subject, we endorse the enclosed findings of the Operational Test Agency Technical Directors and the Science Advisor for Operational Test and Evaluation.

Dr. Charles E. McQueary Director, Operational Test & Evaluation	Stephen T. Sargeant, Major General, USAF Commander, AFOTEC
SES Roger A. Nadeau, Major General, USA Commander, ATEC	David A. Dunaway, Rear Admiral, USN Commander, OPTEVFOR
David L. Reeves, Colonel, USMC Director, MCOTEA	Ronald C. Stephens, Colonel, USA Commander, JITC

Enclosure: Design of Experiments (DOE) in Test and Evaluation

Design of Experiments (DOE) in Test and Evaluation

At the request of the Service Operational Test Agency (OTA) Commanders, DOT&E hosted a meeting of OTA technical and executive agents on February 20, 2009 to consider a common approach to utilizing DOE in operational test and evaluation endeavors. Representatives from ATEC, OPTEVFOR, AFOTEC, JTIC, DOT&E and two experts in DOE from the National Institute of Standards and Technology (NIST) met to discuss the applicability of DOE principles to support test and evaluation efforts.

This group endorses the use of DOE as a discipline to improve the planning, execution, analysis, and reporting of integrated testing. DOE offers a systematic, rigorous, data-based approach to test and evaluation. DOE is appropriate for serious consideration in every case when applied in a testing program. A program applying DOE involves:

- Starting early in the acquisition process with a team of subject matter experts who can identify operational conditions (what they consider the driving factors in the successful performance of the system and the levels of each factor that should be considered)
- Forming a team that must include representation for all testing (Contractor Testing, Government Developmental Testing, Operational Testing), an expert in test design, including DOE, and approval authorities such as DOT&E
- Developing the master plan for the complete test program, the resources needed, and the plan for early tests (even component tests) and use the results of early tests to plan further testing
- Focusing the testing strategy to assure each stage of testing addresses all important parameters, to preclude compartmentalization of specific parameters into separate tests.
- Iterating planning and testing correctly to produce an understanding of the driving factors of system performance and the levels that need to be tested to have an adequate IOT&E that confirms performance.
- Accumulating evidence that the system performs across its operational envelope before and during IOT&E
- Applying DOE as a key ingredient in the formulation of meaningful integrated testing.

Experimental design further provides a valuable tool to identify and mitigate risk in all test activities. It offers a framework from which test agencies may make well-informed decisions on resource allocation and scope of testing required for an adequate test. A DOE-based test approach will not necessarily reduce the scope of resources for adequate testing.

Successful use of DOE will require a cadre of personnel within each OTA organization with the professional knowledge and expertise in applying these methodologies to military test activities. Utilizing the discipline of DOE in all phases of program testing from initial developmental efforts through initial and follow-on operational test endeavors affords the opportunity for rigorous systematic improvement in test processes.

LIST OF REFERENCES

Assistant Secretary of the Air Force for Acquisition (SAF/AQ). (2009). *United States Air Force early systems engineering guidebook, version 1.* Washington, D.C. United States Air Force.

Atkinson, R. (2007, September 30). Left of boom: The struggle to defeat roadside bombs. *The Washington Post* online*,* http://www.washingtonpost.com/wp-srv/world/specials/leftofboom/index.html.

Byers, K. (2010). *Situational awareness for surveillance and interdiction operations (SASIO): Tactical installation operations* (Master's thesis)*.* Monterey, CA: Naval Postgraduate School.

Chaudhary, R. I. (2000, March-April). Reliability Means Performance: Three Recommendations for Program Managers. *Program Manager (PM): A Bimonthly Magazine of the Defense Acquisition University*, 24–28.

Chief of Naval Air Forces (CNAF). (2010). *Naval aviation vision, January 2010.* San Diego, CA: Chief of Naval Air Forces.

Chief of Naval Operations. (2002, May 28). Navy Modeling and Simulation (M&S) Management. *OPNAVINST 5200.34*. Washington, D.C.

Davis, P. K. (2002). *Analytic architecture for capabilities-based planning, mission-system analysis, and transformation.* Santa Monica, CA: RAND National Defense Research Institute.

Defense Science Board Task Force. (2008). *Report of the Defense Science Board Task Force on Developmental Test & Evaluation.* Washington, D.C.: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics.

Department of the Air Force. (2010, July 12). Operational capability requirements development. *Air Force Instruction 10-601*. Washington, D.C.

Deputy Under-Secretary of Defense (Acquisition, Technology, and Logistics). (2007). *Report to Congress on Policies and Practices for Test and Evaluation.* Washington, D.C.: Department of Defense.

Developmental Test and Evaluation (DT&E). (n.d.). Retrieved July 31, 2011, from MDA Glossary: http://www.expertglossary.com/definition/

Devore, J. L. (2009). *Probability and statistics for engineering and the sciences* (7th ed.). Belmont, CA: Brooks/Cole.

Gelman, A. (2005, February 25). *Analysis of variance.* Retrieved July 27, 2011, from http://www.stat.columbia.edu/~gelman/research/unpublished/econanova.pdf

Gilmore, J. M. (2010, October 19). *Guidance on the use of design of experiments (DOE) in operational test and evaluation*. Washington, D.C: Director, Operational Test and Evaluation Command.

Hutto, G. T., & Higdon, J. M. (2009, February). Survey of Design of Experiments (DOE) projects in Developmental Test CY07-08. In *American Institute of Aeronautics and Astronautics 2009-1706.* Paper presented at US Air Force T&E Days, Albuquerque, NM.

Miller, T. H. (2010, July-August). Does MRAP Provide a Model for Acquisition Reform? *Defense AT&L*, pp. 16–20.

Montgomery, D. C. (2009). *Design and analysis of experiments* (7th ed.). New York: John Wiley and Sons, Inc.

Montgomery, D., Peck, E. A., & Vining, G. G. (2006). *Introduction to linear regression analysis.* Hoboken, NJ: John Wiley & Sons, Inc.

Muratore, M. (2010). *Effective teaming of airborne and ground assets for surveillance and interdiction* (Master's thesis)*.* Naval Postgraduate School.

NAVAIRSYSCOM (AIR-5.0). (2011). *Capabilities based test and evaluation working group charter.* Paxtuxent River, MD: NAVAIRSYSCOM.

Office of the Secretary of Defense. (2008, April 25). *Memorandum for component acquisition executives, subject: definition of integrated testing*. Washington, D.C. Retrived 27 May 2011 from http://www.acq.osd.mil/dte/docs/SecDefMemo-Definition-of-Integrated-Testing-25Apr08.pdf

Operational Test Agency Directors & Science Advisor for Operational Test and Evaluation. (2009, February 20). *Design of Experiments (DOE) in Test and Evaluation.* Retrieved July 20, 2011, from Acquisition Community Connection: https://acc.dau.mil/CommunityBrowser.aspx?id=312213

Operational Test and Evaluation of Defense Acquisition Programs, 10 U.S.C § 2399. (2002, January 24). Retrieved July 21, 2011, from GPO Access: http://www.gpoaccess.gov/uscode/index.html

Operational Test and Evaluation (OT&E). (n.d.). Retrieved July 31, 2011, from MDA Glossary: http://www.expertglossary.com/definition/

OPNAV charter (N88). (2009). Naval Aviation Enterprise Capabilities Based Assessment Integrated Process (NCIP) steering group charter. Washington, D.C.: OPNAV (N88).

Phantom Works Communications. (2011, June 27). *Boeing mediaroom.* Retrieved July 14, 2011, from http://boeing.mediaroom.com/index.php?s=43&item=1817

Popper, S. B. (2004, July 21-22). *System-of-systems symposium: Report on a summer conversation.* Retrieved July 13, 2011, from http://www.potomacinstitute.com/academiccen/SoS%20Summer%20Conversation%20report.pdf

SAS Institute Inc. (2011). *JMP design of experiments.* Retrieved from http://www.jmp.com/software/pdf/103044_doe.pdf

Sebolka, J., Grow, D., & Tye, B. (2008, March 29). Testing and training 2020: From stovepipes to collaborative enterprises. *ITEA Journal 2008*, 29: 45–50.

Simpson, J., Hutto, G., & Sewell, A. (2011, May). *Embedding DOE in military testing: one organization's roadmap.* Retrieved July 19, 2011, from http://www.stcnet.com/meetings/2011_NASA_SES/images/presentations/S4A%20-%20Simpson%20-%20Engineering%20Test%20Science%20for%20the%20Military.pdf

Standard, T. (2010, November 10). *Naval aviation enterprise capabilities based assessment integrated process (NCIP).* Patuxent River, MD: Naval Air Systems Command.

Standard, T. (2011, March). *Capabilities based test & evaluation: Delivery of integrated warfighting capabilities.* Briefing to CBTE Working Group, Naval Air Systems Command, Integrated Systems Evaluation, Experimentation and Test, Paxtuxent River.

Streilein, J. J. (2009, March 30). Test and Evaluation of highly complex systems. *ITEA Journal 2009*, 30: 3–6.

Weisgerber, M. (2011, August 19). Collaboration Vital to Acquisition: U.S. CNO. Retrieved August 30, 2011, from Defense News: http://www.defensenews.com/story.php?i=7432668&C=POL&S=TOP

Wu, C. J., & Hamada, M. (2000). *Experiments: Planning, analysis, and parameter design optimization.* New York: John Wiley & Sons, Inc.

Ziegler, B. P., Fulton, D., Hammonds, P., & Nutaro, J. (2005). *Framework for M&S-based system development and testing in a net-centric environment.* Fort Huachuca: Northrop Grumman Information Technology Team Joint Interoperability Test Command.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Malcolm P. Taylor
   Department of Defense, OPNAV, N00X
   Washington, D.C.

4. Leslie Taylor
   Naval Air Systems Command, Air 5.1
   Patuxent River, Maryland

5. Todd Standard
   Naval Air Systems Command, Air 5.1E
   Patuxent River, Maryland

6. Edward Romero
   Naval Air Systems Command, Air 5.1.1
   Pt. Mugu, Ventura County, California

7. Paolo Pringle
   Naval Air Systems Command, Battlespace Engagement 5.1.8.1.0.0E
   Pt. Mugu, Ventura County, California

8. Catherine Warner
   Science Advisor to Director, Operational Test & Evaluation (DOT&E)
   Washington, D.C.

9. Rachel Silvestrini
   Naval Postgraduate School, Operations Research Department
   Monterey, California

10. Harrison Schramm
    Naval Postgraduate School, Operations Research Department
    Monterey, California

11. Tim Chung
    Naval Postgraduate School, Systems Engineering Department
    Monterey, California